

UNIVERSIDAD  
COMPLUTENSE  
DE MADRID



# Ciclo de vida de los datos en Google Cloud

Autor: Pedro Pablo Malagón Amor

Actualizado Enero 2022

## Ciclo de vida de los datos en Google Cloud Platform

## **Ciclo de vida de los datos en Google Cloud Platform**













Vamos a describir los servicios de Google Cloud Platform, que puedes usar para administrar datos a lo largo de todo su ciclo de vida, desde la adquisición inicial hasta la visualización final.

Aprenderás sobre las características y la funcionalidad de cada servicio para que puedas tomar una decisión sobre qué servicios se adaptan mejor a tu carga de trabajo.

El ciclo de vida de los datos tiene cuatro pasos:

1. **Ingestar:** la primera etapa consiste en extraer los datos en bruto, como por ejemplo la transmisión de datos desde dispositivos, datos locales, registros de aplicaciones o análisis y eventos del usuario de una aplicación móvil.
2. **Almacenar:** después de recuperar los datos, es necesario almacenarlos en un formato duradero y de fácil acceso.
3. **Procesar y analizar:** en esta etapa, los datos se transforman de en información procesable.
4. **Explorar y visualizar:** la etapa final consiste en convertir los resultados del análisis en un formato que sea fácil de extraer y compartir.

En cada etapa, Google Cloud Platform proporciona múltiples servicios para administrar tus datos. Esto significa que puedes seleccionar un conjunto de servicios adaptados a tus datos y flujo de trabajo.












Ingest	Store	Process & Analyze	Explore & Visualize
 App Engine  Compute Engine  Container Engine  Cloud Pub/Sub  Stackdriver Logging  Cloud Transfer Service  Transfer Appliance	 Cloud Storage  Cloud SQL  Cloud Datastore  Cloud Bigtable  BigQuery  Cloud Storage for Firebase  Cloud Firestore  Cloud Spanner	 Cloud Dataflow  Cloud Dataproc  BigQuery  Cloud ML  Cloud Vision API  Cloud Speech API  Translate API  Cloud Natural Lang API  Cloud Dataprep  Cloud Video Intelligence API	 Cloud Datalab  Google Data Studio  Google Sheets

## Ingesta

Hay una serie de enfoques que puede tomar para recopilar datos en bruto, en función del tamaño, la fuente y la latencia de los datos.

- **Aplicación:** los datos de los eventos de una aplicación, como los archivos de registro o los eventos del usuario, normalmente se recopilan en un modelo push, donde la aplicación llama a una API para enviar los datos al almacenamiento.
- **Streaming en Transmisión continua:** los datos consisten en un flujo continuo de mensajes pequeños.
- **Batch :** Los grandes volúmenes de datos se almacenan en un conjunto de archivos que se transfieren al almacenamiento de forma masiva.

El siguiente cuadro muestra cómo los servicios de Google Cloud Platform se asignan a las cargas de trabajo de aplicación, streaming y batch.

Applications	Streaming	Batch
 Stackdriver Logging  Cloud Pub/Sub  Cloud SQL  Cloud Datastore  Cloud Bigtable  Cloud Firestore  Cloud Spanner	 Cloud Pub/Sub	 Cloud Storage  Cloud Transfer Service  Transfer Appliance

El modelo de transferencia de datos que elijas depende de su carga de trabajo, y cada modelo tiene diferentes requisitos de infraestructura.

### Ingestar datos de la aplicación

Las aplicaciones y servicios generan una cantidad significativa de datos. Esto incluye datos tales como registros de eventos de la aplicación, datos de seguimiento de clics, interacciones de redes sociales y transacciones de comercio electrónico.

La recopilación y el análisis de estos datos basados en eventos pueden revelar las tendencias de los usuarios y proporcionar información valiosa para el negocio.

Google Cloud Platform proporciona una variedad de servicios que puede usar para alojar aplicaciones, desde las máquinas virtuales de Compute Engine hasta la plataforma administrada de App Engine y la administración de contenedores de Kubernetes Engine.

Cuando alojas tus aplicaciones en Google Cloud Platform, obtienes acceso a herramientas y procesos integrados que facilitan el envío de tus datos al ecosistema de servicios de administración de datos de Google Cloud Platform.

### Consideremos los siguientes ejemplos:

- Escribir datos en un archivo: una aplicación genera archivos CSV por lotes en el almacén de objetos de Cloud Storage. A partir de ahí, la función de importación de BigQuery, puede extraer los datos para su análisis y consulta.

- Escribir datos en una base de datos: una aplicación escribe datos en una de las bases de datos que proporciona Google Cloud Platform, como MySQL o Postgres gestionados de Google Cloud SQL o las bases de datos NoSQL proporcionadas por Cloud Datastore y Google Cloud Bigtable.
- Transmisión de datos como mensajes: una aplicación transmite datos a Cloud Pub / Sub, un servicio de mensajería en tiempo real. Una segunda aplicación, suscrita a los mensajes, puede transferir los datos al almacenamiento o procesarlos inmediatamente en situaciones como la detección de fraude.

### **Stackdriver Logging: gestión centralizada de registros**

Stackdriver Logging es un servicio de administración de registros centralizado que recopila datos de registro de aplicaciones que se ejecutan en Google Cloud Platform y otras plataformas de nube públicas y privadas.

La exportación de datos recopilados por Stackdriver Logging es fácil de enviar a las herramientas integradas que envían los datos a Cloud Storage, Cloud Pub / Sub y BigQuery.

Muchos servicios de Google Cloud Platform registran automáticamente datos de registro en Stackdriver Logging. Por ejemplo, las aplicaciones que se ejecutan en App Engine registran automáticamente los detalles de cada solicitud y respuesta a Stackdriver Logging.

También puede escribir mensajes de registro personalizados en stdout y stderr, que Stackdriver Logging recopila automáticamente y muestra en Logs Viewer.

Stackdriver Logging proporciona un agente de registro, basado en fluentd, que puede ejecutar en instancias de máquinas virtuales (VM) alojadas en Compute Engine, así como en clústeres de contenedores administrados por Kubernetes Engine. El agente transmite datos de registro desde aplicaciones de terceros comunes y software del sistema a Stackdriver Logging.

### **Ingestión de datos de streaming**

Los datos de streaming se entregan de manera asíncrona, sin esperar una respuesta, y los mensajes individuales son de pequeño tamaño.

Comúnmente, la transmisión de datos se utiliza para la telemetría, recopilando datos de dispositivos geográficamente dispersos. Los datos de transmisión se pueden usar para disparar activadores de eventos, realizar análisis de sesiones complejos y como entrada para tareas de aprendizaje automático.

Los usos comunes de la transmisión de datos incluyen:

- Datos de telemetría: los dispositivos de Internet de las cosas (IoT) son dispositivos conectados a la red que recopilan datos del entorno a través de sensores. Aunque

cada dispositivo puede enviar una gran cantidad de datos por minuto, cuando multiplicas esos datos por una gran cantidad de dispositivos, rápidamente necesitas aplicar estrategias y patrones de big data.

- Eventos de usuario y análisis: una aplicación móvil puede registrar eventos de usuario cuando el usuario abre la aplicación y siempre que ocurre un error o bloqueo. El conjunto de estos datos, en todos los dispositivos móviles donde se instala la aplicación, puede proporcionar información valiosa sobre el uso, las métricas y la calidad del código.

### **Cloud Pub / Sub: mensajería en tiempo real**

Cloud Pub / Sub es un servicio de mensajería en tiempo real que le permite enviar y recibir mensajes entre aplicaciones. Uno de los principales casos de uso para la mensajería entre aplicaciones es la ingesta de datos de eventos de transmisión.

Con la transmisión de datos, Cloud Pub / Sub gestiona automáticamente los detalles de fragmentación, replicación, equilibrio de carga y partición de las secuencias de datos entrantes.

La mayoría de los datos de transmisión son generados por usuarios o sistemas distribuidos en todo el mundo. Cloud Pub / Sub tiene puntos finales globales y aprovecha el equilibrador de carga front-end global de Google para admitir la ingestión de datos en todas las regiones de Google Cloud Platform, con una latencia mínima.

Además, Cloud Pub / Sub escala de forma rápida y automática para satisfacer la demanda, sin que el desarrollador pre-aprovise los recursos del sistema.

Los temas son cómo Cloud Pub / Sub organiza los flujos de mensajes.

Las aplicaciones que transmiten datos a Cloud Pub / Sub se dirigen a un tema específico. Cuando recibe cada mensaje, Cloud Pub / Sub adjunta un identificador único e indicación de fecha y hora.

Una vez que se han ingerido los datos, una o más aplicaciones pueden recuperar los mensajes mediante una suscripción temática.

Esto se puede hacer en un modelo pull o push.

En una suscripción push, el servidor Pub / Sub envía una solicitud a la aplicación del suscriptor en un punto final URL preconfigurado.

En el modelo de extracción, el suscriptor solicita mensajes del servidor y acusa recibo. Cloud Pub / Sub garantiza la entrega de mensajes al menos una vez por suscriptor.

Cloud Pub / Sub no proporciona garantías sobre el orden de entrega del mensaje. El ordenamiento estricto de los mensajes se puede lograr con el almacenamiento en búffer, a menudo usando Cloud Dataflow.

Un uso común de Cloud Pub / Sub es mover datos de transmisión en Cloud Dataflow para el procesamiento en tiempo real, por tiempo real del evento. Cuando se procesa, puede mover los datos a un servicio de almacenamiento persistente, como Cloud Datastore y BigQuery, que admiten consultas ordenadas por marcas de tiempo de la aplicación.

### **Ingerir datos en bruto**

Los datos masivos constan de grandes conjuntos de datos donde la ingestión requiere un gran ancho de banda agregado entre un pequeño número de fuentes y el objetivo.

Los datos se pueden almacenar en archivos, como archivos CSV, JSON, Avro o Parquet, o en una base de datos relacional o NoSQL.

Considera los siguientes ejemplos:

- Carga de trabajo científica: los datos genéticos almacenados en archivos de texto de formato (VCF) se cargan en Cloud Storage para su posterior importación a Genomics.
- Migración a la nube: mover datos almacenados en una base de datos Oracle local a una base de datos Google Cloud SQL totalmente administrada.
- Copia de seguridad de datos: replicación de datos almacenados en un depósito de AWS en Cloud Storage mediante Cloud Storage Transfer Service.
- Importación de datos heredados: Copia de diez años de datos de registro del sitio web en BigQuery para el análisis de tendencias a largo plazo.

Google Cloud Platform y las empresas asociadas ofrecen una variedad de herramientas que puede usar para cargar grandes conjuntos de datos en Google Cloud Platform.

### **Servicio de transferencia de almacenamiento en la nube: transferencia de archivos administrada**

Cloud Storage Transfer Service gestiona la transferencia de datos a un Bucket de Cloud Storage. La fuente de datos puede ser un depósito AWS S3, una URL accesible en la web u otro Bucket de Cloud Storage.

Cloud Storage Transfer Service está destinado a la transferencia masiva y está optimizado para volúmenes de datos superiores a 1TB.

Realizar una copia de seguridad de datos es un uso muy común del Servicio de Cloud Storage Transfer. Puedes hacer una copia de seguridad de los datos de otros proveedores de almacenamiento en un Bucket de Cloud Storage. O bien, puedes mover datos entre



sectores de Cloud Storage, como archivar datos de un bucket de almacenamiento multirregional a un segmento de almacenamiento Nearline para reducir los costes de almacenamiento.

Cloud Storage Transfer Service admite transferencias únicas o recurrentes. Proporciona filtros avanzados basados en las fechas de creación de archivos, filtros de nombre de archivo y las horas del día en que prefiere importar datos. También es compatible con la eliminación de los datos de origen después de que se haya copiado.

**Dispositivo de transferencia:  
servidor de almacenamiento de alta capacidad que se puede enviar**

Google Transfer Appliance es un servidor de almacenamiento de gran capacidad que alquila de Google. Lo conectas a tu red, lo cargas con tus datos y lo envías a una instalación de Google, donde los datos se cargan en Cloud Storage.

Google Transfer Appliance viene en varios tamaños, además, dependiendo de la naturaleza de tus datos, es posible que pueda usar deduplicación y compresión para aumentar sustancialmente la capacidad efectiva del dispositivo.

Para determinar cuándo usar Google Transfer Appliance, calcula la cantidad de tiempo necesario para cargar tus datos mediante una conexión de red. Si determinas que tardarías una semana o más, o si tienes más de 60 TB de datos (independientemente de la velocidad de transferencia), podría ser más conveniente transferir tus datos utilizando Google Transfer Appliance.

Transfer Appliance deduplica, comprime y cifra tus datos capturados con un fuerte cifrado AES-256 utilizando una contraseña y frase de contraseña que usted proporciona.

Cuando lees tus datos de Cloud Storage, especifica la misma contraseña y frase de contraseña. Después de cada uso de Transfer Appliance, el dispositivo se limpia de forma segura y se vuelve a crear una imagen para ayudar a evitar que tus datos estén disponibles para el próximo usuario.

**Cloud Storage gsutil: interfaz de línea de comandos**

Cloud Storage proporciona gsutil, una utilidad de línea de comandos que puedes usar para mover datos basados en archivos de cualquier sistema de archivos existente a Cloud Storage.

Escrito en Python, gsutil se ejecuta en sistemas Linux, macOS y Windows. Además de mover datos a Cloud Storage, puede usar gsutil para crear y administrar depósitos de Cloud Storage, editar derechos de acceso de objetos y copiar objetos de Cloud Storage.

**Almacenamiento en la nube offline: importación / exportación**

Offline Media Import / Export es una solución de terceros que puede usar para cargar datos en Cloud Storage enviando tus medios físicos, como unidades de disco duro, cintas y unidades flash USB, a un proveedor de servicios tercero que carga datos en su nombre.

La importación / exportación de medios sin conexión es útil si está limitado a una conexión a Internet lenta, poco fiable o muy cara.

### **Herramientas de migración de bases de datos**

Si tus datos de origen están almacenados en una base de datos, ya sea localmente o alojados por otro proveedor de la nube, hay varias aplicaciones de terceros que puede usar para migrar tus datos en bruto, a Google Cloud Platform. Estas aplicaciones a menudo se ubican en el mismo entorno que los sistemas de origen y proporcionan transferencias únicas y continuas. Aplicaciones como Talend e Informatica proporcionan capacidades de extracción-transformación-carga (ETL) con soporte integrado para Google Cloud Platform.

Google Cloud Platform tiene varias bases de datos de destino adecuadas para migrar datos desde bases de datos externas.

- Bases de datos relacionales: los datos almacenados en un sistema de administración de bases de datos relacionales (RDBMS) se pueden migrar a Google Cloud SQL y Google Cloud Spanner.
- Data Warehouse: los datos almacenados en un data warehouse se puede mover a BigQuery.
- Bases de datos NoSQL: los datos almacenados en una base de datos NoSQL orientada a columnas, como HBase o Cassandra, se pueden migrar a Google Cloud Bigtable. Los datos almacenados en una base de datos NoSQL, como Couchbase o MongoDB, se pueden migrar a Cloud Datastore.

### **Soluciones de Partners**

Varios Partner de Google Cloud Platform proporcionan soluciones complementarias centradas en el movimiento masivo de datos.

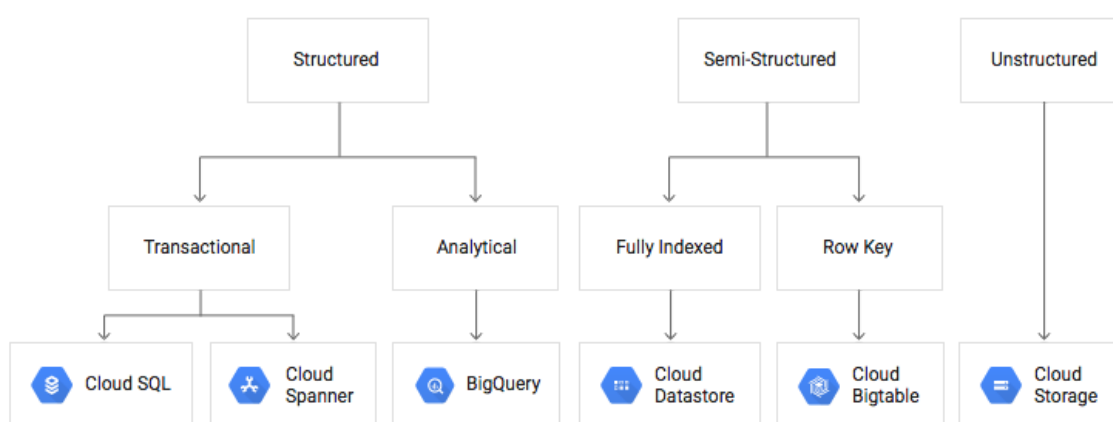
- WANDisco proporciona Google Active Migrator, que automatiza la transferencia de datos desde el almacenamiento local local y de red a los clústeres Cloud Dataproc.
- Tervela ofrece Cloud FastPath, para automatizar la migración de datos y la sincronización del sistema de archivos local con Cloud Storage.

- Tanto Iron Mountain como Prime Focus ofrecen la posibilidad de cargar datos en Cloud Storage desde tus medios físicos, como unidades de disco duro, cintas y unidades flash USB.

### Almacenamiento

Los datos vienen en diferentes formas y tamaños y su estructura depende totalmente de las fuentes de donde se generó y los casos posteriores de uso.

Para las cargas de trabajo de datos y análisis, los datos ingresados se pueden almacenar en una variedad de formatos o ubicaciones.



### Almacenamiento de datos de objetos

Los archivos son un formato común para almacenar datos, especialmente datos masivos. Con Google Cloud Platform puedes cargar tus datos de archivo a Google Cloud Storage, lo que hace que esos datos estén disponibles para una variedad de otros servicios.

### Almacenamiento en la nube: almacenamiento de objetos administrados

Cloud Storage ofrece almacenamiento de objetos duradero y altamente disponible para datos estructurados y no estructurados. Ejemplos de tales datos puede ser archivos de registro, copias de seguridad de bases de datos y archivos de exportación, imágenes y otros archivos binarios. Los archivos en Google Cloud Storage están organizados por proyecto en categorías individuales. Estos depósitos pueden admitir listas de control de acceso (ACL) personalizadas o controles centralizados de gestión de identidades y acceso (IAM).

Google Cloud Storage actúa como una capa de almacenamiento distribuido, accesible para aplicaciones y servicios que se ejecutan en App Engine, Kubernetes Engine o Compute Engine, y a través de otros servicios como Google Cloud Logging.

**Considera los siguientes casos de uso:**

- Copia de seguridad de datos y recuperación de desastres: Google Cloud Storage ofrece almacenamiento altamente durable y seguro para realizar copias de seguridad y archivar tus datos.
- Distribución de contenido: Google Cloud Storage permite el almacenamiento y la entrega de contenido. Por ejemplo, almacenar y entregar archivos multimedia es simple y escalable.
- Almacenamiento de datos de ETL: Google Cloud Dataflow puede acceder a los datos de Google Cloud Storage para su transformación y carga en otros sistemas, como Google Cloud Bigtable o Google Cloud BigQuery.
- Almacenamiento de datos para trabajos de MapReduce: para trabajos de Hadoop y Spark, se puede acceder de forma nativa a los datos de Google Cloud Storage con Google Cloud Dataproc.
- Almacenamiento de datos de consultas: Google Cloud BigQuery tiene la capacidad de importar datos de Google Cloud Storage en conjuntos de datos y tablas, o las consultas se pueden federar a través de los datos existentes sin importar. Para el acceso directo, Google Cloud BigQuery admite de forma nativa la importación de archivos CSV, JSON y Avro desde un depósito de Google Cloud Storage especificado.
- Aprendizaje automático : las API de Machine Learning Google Cloud Platform, como Visión API o Natural Language API, pueden acceder a datos y archivos almacenados directamente en Google Cloud Storage.
- Archivar datos fríos: Nearline Storage y Coldline Storage ofrecen un almacenamiento de baja latencia y menor coste para los objetos a los que planeas acceder menos de una vez al mes o menos de una vez al año, respectivamente.

Google Cloud Storage está disponible en varias clases, según la disponibilidad y el rendimiento requeridos para las aplicaciones y los servicios.

- El almacenamiento multirregional ofrece los niveles más altos de disponibilidad y es apropiado para almacenar datos que requieren acceso altamente redundante y de baja latencia para los datos a los que se accede con frecuencia. Los ejemplos de casos de uso incluyen el contenido del sitio web, las cargas de trabajo de almacenamiento interactivo y los datos que admiten aplicaciones móviles y de juegos.

- Regional Storage ofrece almacenamiento de alto rendimiento dentro de una sola región y es apropiado para almacenar datos utilizados por las instancias de Compute Engine. Ejemplos de casos de uso incluyen cálculos de datos intensivos o procesamiento de big data.
- Nearline Storage es un servicio de almacenamiento de bajo costo y muy duradero para almacenar datos a los que accede menos de una vez al mes. Nearline Storage ofrece un acceso rápido a los datos, en el orden de los tiempos de respuesta por debajo del segundo y es útil para el archivado de datos, la copia de seguridad en línea o casos de uso de recuperación de desastres.
- Coldline Storage ofrece un servicio de almacenamiento de muy bajo costo y muy duradero para almacenar datos a los que tiene la intención de acceder menos de una vez al año. Coldline Storage ofrece acceso rápido a los datos, en el orden de los tiempos de respuesta por debajo del segundo, y es apropiado para el archivo de datos, la copia de seguridad en línea y la recuperación ante desastres.

### **Cloud Storage para Firebase: almacenamiento escalable para desarrolladores de aplicaciones móviles**

Cloud Storage for Firebase es un servicio de almacenamiento de objetos simple y rentable diseñado para escalar con su base de usuarios. Google Cloud Storage for Firebase es una buena opción para almacenar y recuperar activos como imágenes, audio, video y otro contenido generado por el usuario dentro de las aplicaciones móviles y web.

Los SDK de Firebase para almacenamiento en la nube realizan cargas y descargas independientemente de la calidad de la red. Si se interrumpen debido a conexiones deficientes, se reinician donde se detuvieron, ahorrando a los usuarios tiempo y ancho de banda. La integración lista para usar con Firebase Authentication te permite configurar el acceso según el nombre de archivo, el tamaño, el tipo de contenido y otros metadatos.

Cloud Storage for Firebase almacena tus archivos en un bucket de Cloud Storage, lo que te da la flexibilidad de cargar y descargar archivos desde clientes móviles que usan Firebase SDK. También puede realizar el procesamiento del lado del servidor, como el filtrado de imágenes o la transcodificación de video con Google Cloud Platform.

Firebase tiene SDK para iOS, Android, web, C ++ y Unity.

### **Almacenamiento de datos en base de datos**

Google Cloud Platform proporciona una variedad de bases de datos, tanto RDBMS como NoSQL, que puede usar para almacenar tus datos relacionales y no relacionales.

### **Google Cloud SQL: motores administrados MySQL y PostgreSQL**

Google Cloud SQL es un RDBMS completamente administrado y nativo de la nube que ofrece motores MySQL y PostgreSQL con soporte incorporado para la replicación.

Es útil para cargas de trabajo de bases de datos relacionales, transaccionales y de baja latencia. Debido a que está basado en MySQL y PostgreSQL, Google Cloud SQL admite el API estándar para la conectividad.

Google Cloud SQL ofrece copia de seguridad y restauración integradas, alta disponibilidad y réplicas de lectura.

Google Cloud SQL admite cargas de trabajo RDBMS de hasta 10 TB tanto para MySQL como para PostgreSQL. Se puede acceder a Google Cloud SQL desde aplicaciones que se ejecutan en App Engine, Kubernetes Engine o Compute Engine. Debido a que Google Cloud SQL está construido sobre MySQL y PostgreSQL, admite controladores de conexión estándar, marcos de aplicaciones de terceros (como Django y Ruby on Rails) y herramientas de migración populares.

Los datos almacenados en Google Cloud SQL están cifrados tanto en tránsito como en reposo. Las instancias de Google Cloud SQL tienen soporte integrado para el control de acceso, usando firewalls de red para administrar el acceso a la base de datos.

Google Cloud SQL es apropiado para las cargas de trabajo típicas de procesamiento de transacciones en línea (OLTP), tales como:

- Transacciones financieras: el almacenamiento de transacciones financieras requiere la semántica de la base de datos ACID y los datos a menudo se distribuyen en varias tablas, por lo que se requiere un soporte de transacción complejo.
- Credenciales del usuario: el almacenamiento de contraseñas u otros datos seguros requiere un soporte de campo complejo y la aplicación junto con la validación del esquema.
- Pedidos de clientes: los pedidos o facturas suelen incluir datos relacionales altamente normalizados y soporte de transacciones de múltiples mesas al capturar cambios de inventario.

Google Cloud SQL no es un sistema de almacenamiento apropiado para cargas de trabajo de procesamiento analítico en línea (OLAP) o datos que requieren esquemas dinámicos por objeto.

- Si tu carga de trabajo requiere esquemas dinámicos, considera Cloud Datastore.
- Para cargas de trabajo OLAP, considera BigQuery.
- Si tu carga de trabajo requiere esquemas de columna ancha, considera Google Cloud Bigtable.

Para procesamiento descendente y casos de uso analítico, se puede acceder a los datos en Google Cloud SQL desde múltiples herramientas de plataforma.

Puedes usar Google Cloud Dataflow o Google Cloud Dataproc para crear trabajos ETL que extraigan datos de Google Cloud SQL e inserten en otros sistemas de almacenamiento.

### **Google Cloud Bigtable: NoSQL gestionado columnar**

Google Cloud Bigtable es un servicio de base de datos NoSQL administrado y de alto rendimiento diseñado para cargas de trabajo de escala de terabyte a petabyte. Google Cloud Bigtable se basa en la infraestructura de base de datos interna de Bigtable de Google que impulsa la Búsqueda de Google, Google Analytics, Google Maps y Gmail. El servicio proporciona almacenamiento consistente, de baja latencia y de alto rendimiento para datos NoSQL a gran escala. Google Cloud Bigtable está diseñado para cargas de trabajo de aplicación en tiempo real, así como cargas de trabajo analíticas a gran escala.

Los esquemas de Google Cloud Bigtable usan una clave de fila indexada individual asociada con una serie de columnas; los esquemas generalmente se estructuran como “altos” o “anchos” y las consultas se basan en la clave de fila. El estilo del esquema depende de los casos de uso posteriores y es importante considerar la ubicación de los datos y la distribución de lecturas y escrituras para maximizar el rendimiento.

Los esquemas “altos” a menudo se utilizan para almacenar eventos de series de tiempo, datos que están codificados en alguna parte por una marca de tiempo, con relativamente pocas columnas por fila.

Los esquemas “anchos” siguen el enfoque opuesto, un identificador simplista como la clave de fila junto con una gran cantidad de columnas.

Google Cloud Bigtable es ideal para una variedad de cargas de trabajo de alto rendimiento a gran escala, como la tecnología publicitaria o la infraestructura de datos IoT.

Considera los siguientes casos de uso:

- Datos de aplicaciones en tiempo real: se puede acceder a Google Cloud Bigtable desde aplicaciones que se ejecutan en App Engine Flexible, Kubernetes Engine y Compute Engine para cargas de trabajo en vivo en tiempo real.
- Procesamiento de flujo: a medida que Cloud Pub / Sub ingiere los datos, Cloud Dataflow se puede utilizar para transformar y cargar los datos en Google Cloud Bigtable.
- Datos de series de tiempo de IoT: los datos capturados por sensores y transmitidos a Cloud Platform se pueden almacenar utilizando esquemas de series de tiempo en Google Cloud Bigtable.

- Carga de trabajo de Adtech: Google Cloud Bigtable se puede usar para almacenar y rastrear impresiones de anuncios, así como también una fuente para el procesamiento y análisis de seguimiento utilizando Cloud Dataproc y Cloud Dataflow.
- Ingestión de datos: Cloud Dataflow o Cloud Dataproc se pueden usar para transformar y cargar datos de Cloud Storage en Google Cloud Bigtable.
- Cargas de trabajo analíticas: Cloud Dataflow se puede usar para realizar agregaciones complejas directamente a partir de datos almacenados en Google Cloud Bigtable, y Cloud Dataproc se puede usar para ejecutar tareas de procesamiento automático y de procesamiento de Hadoop o Spark.
- Reemplazo de Apache HBase: Google Cloud Bigtable también se puede utilizar como reemplazo directo para los sistemas creados con Apache HBase, una base de datos de código abierto basada en el documento original de Bigtable creado por Google. Google Cloud Bigtable cumple con las API de HBase 1.x, por lo que se puede integrar fácilmente en muchos sistemas de big data existentes. Apache Cassandra utiliza un modelo de datos basado en el que se encuentra en el documento de Bigtable, lo que significa que Google Cloud Bigtable también puede admitir varias cargas de trabajo que aprovechan un esquema y estructura orientados a columnas amplias.

Si bien Google Cloud Bigtable se considera un sistema OLTP, no admite transacciones de varias filas, consultas SQL o combinaciones. Para esos casos de uso, considera Google Cloud SQL o Cloud Datastore.



## **Google Cloud Spanner: base de datos relacional horizontalmente escalable**

Google Cloud Spanner es un servicio de base de datos relacional totalmente administrado para aplicaciones OLTP de misión crítica.

Google Cloud Spanner es escalable horizontalmente y está diseñado para ofrecer una gran consistencia, alta disponibilidad y escala global.

Esta combinación de cualidades lo hace único como un servicio.

Debido a que Google Cloud Spanner es un servicio totalmente administrado, puedes enfocarte en diseñar tu aplicación y no tu infraestructura.

Google Cloud Spanner es una buena opción para los clientes que desean la facilidad de uso y la familiaridad de una base de datos relacional junto con la escalabilidad típicamente asociada con una base de datos NoSQL.

Al igual que las bases de datos relacionales, Google Cloud Spanner admite esquemas, transacciones ACID y consultas SQL (ANSI 2011).

Al igual que muchas bases de datos NoSQL, Google Cloud Spanner se escala horizontalmente dentro de las regiones, pero también puede escalar regiones para cargas de trabajo que tienen requisitos de disponibilidad más estrictos.

Google Cloud Spanner también realiza fragmentación automática mientras sirve datos con latencias de milisegundos de un solo dígito.

Las características de seguridad en Google Cloud Spanner incluyen cifrado de capa de datos, registro de auditoría e integración con IAM.

Los casos de uso típicos de Google Cloud Spanner son:

- **Servicios financieros:** las cargas de trabajo de los servicios financieros requieren una gran coherencia en todas las operaciones de lectura / escritura. Google Cloud Spanner proporciona esta coherencia sin sacrificar la alta disponibilidad.
- **Ad tech-Latency** es una consideración clave en el espacio de tecnología publicitaria. Google Cloud Spanner facilita la consulta de baja latencia sin comprometer la escala o la disponibilidad.
- **Cadena de suministro minorista y global:** la necesidad de una escala global puede obligar a los expertos en cadenas de suministro a hacer concesiones entre la coherencia y los costes de mantenimiento. Google Cloud Spanner ofrece replicación automática, global y sincrónica con baja latencia, lo que significa que los datos son siempre consistentes y de alta disponibilidad.

### **Cloud Firestore: base de datos NoSQL flexible y escalable**

Cloud Firestore es una base de datos que almacena datos JSON.

Los datos JSON se pueden sincronizar en tiempo real con clientes conectados en diferentes plataformas, incluidos iOS, Android, JavaScript, dispositivos IoT y aplicaciones de escritorio. Si un cliente no tiene conectividad de red, el SDK de Cloud Firestore permite que su aplicación conserve los datos en un disco local. Después de restablecer la conectividad, el dispositivo cliente se sincroniza con el estado actual del servidor.

Cloud Firestore proporciona un lenguaje de reglas flexible basado en expresiones, Cloud Firestore Security Rules, que se integra con Firebase Authentication para que pueda definir quién tiene acceso a qué datos.

Cloud Firestore es una base de datos NoSQL con una API que puede usar para crear una experiencia en tiempo real que sirva a millones de usuarios sin comprometer la capacidad de respuesta. Para facilitar este nivel de escala y capacidad de respuesta, es importante estructurar tus datos de manera apropiada.

Cloud Firestore tiene SDK para iOS, Android, web, C++ y Unity.

Aquí hay un par de casos de uso para Cloud Firestore:

- Chat y redes sociales: almacena y recupera imágenes, audio, video y otro contenido generado por los usuarios.
- Juegos móviles: haz un seguimiento del progreso del juego y las estadísticas en todos los dispositivos y plataformas de dispositivos.

### **Ecosistema de Bases de datos**

Además de los servicios de base de datos proporcionados por Google Cloud Platform, tu puedes desplegar tu propio software de base de datos en máquinas virtuales de Google Compute Engine, las cuales son VMs de alto rendimiento y con almacenamiento persistente altamente escalable.

Los RDBMS tradicionales como SAP HANA o Microsoft SQL Server son compatibles con Google Cloud Platform. Los sistemas de bases de datos NoSQL como MongoDB y Cassandra también son compatibles con configuraciones de alto rendimiento.

Con Cloud Launcher puede implementar rápidamente muchos tipos de bases de datos en Google Cloud Platform usando imágenes preconstruidas, almacenamiento y configuraciones de red. Los recursos de implementación, como las instancias de Google Compute Engine, los discos persistentes y las configuraciones de red, se pueden gestionar de forma directa y sencilla para diferentes cargas de trabajo o casos de uso.

### **Almacenamiento de datos en data warehouse**

Un data warehouse almacena grandes cantidades de datos para consulta y análisis en lugar de procesamiento transaccional. Para las cargas de trabajo de data warehouse, Google Cloud Platform proporciona BigQuery.

### **Google Cloud BigQuery: data warehouse administrados**

Para los datos ingestados que se analizarán en última instancia dentro de BigQuery, puedes almacenar datos directamente en BigQuery, pasando por alto otros medios de almacenamiento.

BigQuery admite la carga de datos a través de la interfaz web, las herramientas de línea de comandos y las llamadas a la API REST.

Al cargar datos en bruto, los datos deben estar en forma de archivos CSV, JSON o Avro.

A continuación, puede usar la interfaz web de BigQuery, las herramientas de línea de comandos o las llamadas a la API REST para cargar datos de estos formatos de archivo en tablas de BigQuery.

Para la transmisión de datos, puedes usar Cloud Pub / Sub y Google Cloud Dataflow en combinación para procesar las transmisiones entrantes y almacenar los datos resultantes en BigQuery.

En algunas cargas de trabajo, sin embargo, puede ser apropiado transmitir datos directamente a BigQuery sin procesamiento adicional.

También puedes crear aplicaciones personalizadas, que se ejecutan en Google Cloud Platform o en infraestructura local, que leen desde orígenes de datos con esquemas y filas definidos. La aplicación personalizada puede transmitir esos datos a tablas de BigQuery utilizando los SDK de Google Cloud o las llamadas a API REST directas.

### **Procesar y analizar**




Para derivar el valor comercial y las perspectivas de los datos, debes transformarlos y analizarlos. Esto requiere un marco de procesamiento que pueda analizar los datos directamente o preparar los datos para el análisis posterior, así como las herramientas para analizar y comprender los resultados del procesamiento.

- **Procesamiento:** los datos de los sistemas fuente se limpian, normalizan y procesan en varias máquinas y se almacenan en sistemas analíticos.
- **Análisis:** los datos procesados se almacenan en sistemas que permiten consultas y exploraciones ad-hoc.
- **Comprensión:** en función de los resultados analíticos, los datos se utilizan para capacitar y probar modelos automáticos de aprendizaje automático.

Google Cloud Platform proporciona servicios para procesar datos a gran escala, analizar y consultar big data y comprender datos a través del aprendizaje automático.

### Procesando datos a gran escala

El procesamiento de datos a gran escala generalmente implica leer datos de sistemas fuente como Cloud Storage, Google Cloud Bigtable o Google Cloud SQL, y luego realizar normalizaciones complejas o agregaciones de esos datos. En muchos casos, los datos son demasiado grandes para caber en una sola máquina, por lo tanto se utilizan frameworks para administrar clusters de cómputo distribuidos y para proporcionar herramientas de software que ayudan al procesamiento.

 Cloud Dataproc	 Cloud Dataflow	 Cloud Dataprep
<ul style="list-style-type: none"><li>• Existing Hadoop/Spark Applications</li><li>• Machine Learning / Data Science Ecosystem</li><li>• Tunable Cluster Parameters</li></ul>	<ul style="list-style-type: none"><li>• New Data Processing Pipelines</li><li>• Unified Streaming &amp; Batch</li><li>• Fully-Managed, No-Ops</li></ul>	<ul style="list-style-type: none"><li>• UI-Driven Data Preparation</li><li>• Scales On-Demand</li><li>• Fully-Managed, No-Ops</li></ul>

### **Cloud Dataproc: Apache Hadoop y Apache Spark administrados**

La capacidad de tratar con conjuntos de datos extremadamente grandes ha evolucionado desde que Google publicó por primera vez el documento MapReduce en 2004. Muchas organizaciones ahora cargan y almacenan datos en Hadoop Distributed File System (HDFS) y ejecutan agregaciones, informes o transformaciones periódicas utilizando herramientas tradicionales de lotes como Hive o Pig. Hadoop tiene un gran ecosistema para admitir actividades como el aprendizaje automático usando Mahout, la ingestión de registros utilizando Flume y las estadísticas con R, y más. Los resultados de este procesamiento de datos basado en Hadoop son críticos para el negocio. No es un ejercicio trivial para una organización dependiente de estos procesos migrarlos a un nuevo marco.

Spark ha ganado popularidad en los últimos años como una alternativa extremadamente rápida y simple a Hadoop MapReduce. El rendimiento de Spark es generalmente considerablemente más rápido que Hadoop MapReduce. Spark logra esto mediante la distribución de conjuntos de datos y computación en la memoria a través de un clúster. Además de los aumentos de velocidad, esta distribución le da a Spark la capacidad de manejar datos de transmisión mediante Spark Streaming, así como análisis, transformaciones y agregaciones de lotes tradicionales mediante el uso de SQL mediante Spark SQL y una API simple. La comunidad Spark es muy activa con varias bibliotecas populares, incluyendo MLlib, que se puede utilizar para el aprendizaje automático.

Sin embargo, ejecutar Spark o Hadoop en una escala cada vez mayor crea complejidad operativa y gastos generales, así como un costo fijo continuo y creciente. Incluso si solo se necesita un clúster en intervalos discretos, aún así se termina pagando el costo de un clúster persistente. Con Cloud Dataproc, puedes migrar tus implementaciones existentes de Hadoop o Spark a un servicio completamente administrado que automatiza la creación de clústeres, simplifica la configuración y administración de su clúster, cuenta con informes integrados de monitorización y utilización, y se puede apagar cuando no esté en uso.

El inicio de un nuevo clúster de Cloud Dataproc tarda, en promedio, 90 segundos, lo que facilita la creación de un clúster de 10 nodos o incluso un clúster de 1000 nodos. Esto reduce la sobrecarga operacional y de costes de la administración de una implementación de Spark o Hadoop, a la vez que proporciona la familiaridad y consistencia de cualquiera de los marcos. Cloud Dataproc proporciona la facilidad y la flexibilidad para activar los clústeres Spark o Hadoop cuando se necesitan y para finalizar los clústeres cuando ya no se necesitan.

Considera los siguientes casos de uso:

- **Procesamiento de registro:** con modificaciones mínimas, puedes procesar grandes cantidades de datos de registro de texto por día desde varias fuentes usando MapReduce existente.

- Informes: agrega datos en informes y almacena los datos en BigQuery. Luego puedes enviar los datos agregados a las aplicaciones y realizar análisis.
- Clústeres de Spark a petición: inicia rápidamente clústeres ad-hoc para analizar los datos que se almacenan en el almacenamiento de blobs utilizando Spark (Spark SQL, PySpark, Shell Spark).
- Aprendizaje automático: utiliza las bibliotecas de aprendizaje automático Spark (MLlib), que están preinstaladas en el clúster, para personalizar y ejecutar algoritmos de clasificación.

Cloud Dataproc también simplifica las actividades operativas, como la instalación de software o el cambio de tamaño de un clúster. Con Cloud Dataproc, puedes leer datos de forma nativa y escribir resultados en Cloud Storage, Google Cloud Bigtable o BigQuery, o en el almacenamiento HDFS proporcionado por el clúster. Con Cloud Storage, Cloud Dataproc se beneficia de un acceso más rápido a los datos y la capacidad de tener muchos clústeres que operan de manera uniforme en conjuntos de datos sin movimiento de datos, además de eliminar la necesidad de centrarse en la replicación de datos. Esta capacidad de almacenar y controlar los datos del punto de control externamente hace posible que trates los clústeres de Dataproc como recursos efímeros con persistencia externa, que se pueden iniciar, consumir y finalizar según sea necesario.

### **Cloud Dataflow: procesamiento de lotes y flujo totalmente administrado y sin servidor**

Poder analizar los datos de transmisión ha transformado la manera en que las organizaciones hacen negocios y cómo responden en tiempo real. Sin embargo, tener que mantener diferentes marcos de procesamiento para tratar el análisis por lotes y de transmisión aumenta la complejidad al necesitar dos canales diferentes. Y perder tiempo optimizando la utilización y los recursos del clúster, como lo hace con Spark y Hadoop, distrae del objetivo básico de filtrar, agregar y transformar tus datos.

Cloud Dataflow se diseñó para simplificar Big Data tanto para cargas de trabajo continuas como por lotes. Lo hace unificando el modelo de programación y el modelo de ejecución. En lugar de tener que especificar un tamaño de clúster y administrar la capacidad, Cloud Dataflow es un servicio administrado en el que se crean recursos a pedido, se escalan automáticamente y se paraleliza. Como un verdadero servicio de operación cero, los trabajadores son agregados o eliminados en función de las demandas del trabajo. Cloud Dataflow también se ocupa del problema común de los trabajadores rezagados que se encuentran en los sistemas distribuidos al monitorear, identificar y reprogramar constantemente el trabajo, incluidas las divisiones, a los trabajadores inactivos en todo el clúster.

Considera los siguientes casos de uso:

- MapReduce replacement: procesa cargas de trabajo paralelas donde los paradigmas de procesamiento que no son de MapReduce han llevado a una complejidad o frustración operativa.
- Análisis de usuarios: analiza datos de alto volumen de comportamiento del usuario, como eventos en el juego, datos de transmisión de clics y datos de ventas minoristas.
- Ciencia de datos: procesa grandes cantidades de datos para hacer descubrimientos y predicciones científicas, como la genómica, el clima y los datos financieros.
- ETL: ingiere, transforma y carga datos en un almacén de datos, como BigQuery.
- Procesamiento de registro: procesa el procesamiento de datos de registro de eventos continuo para crear cuadros de mando en tiempo real, métricas de aplicaciones y alertas.

El SDK de Dataflow también se ha lanzado como el proyecto de código abierto Apache Beam que admite la ejecución en Apache Spark y Apache Flink. Debido a su escalado automático y facilidad de implementación, Cloud Dataflow es una ubicación ideal para ejecutar flujos de trabajo de flujo de datos / Apache.

## **Cloud Dataprep: Exploración, limpieza y procesamiento de datos visuales**

Cloud Dataprep es un servicio para explorar visualmente, limpiar y preparar datos para el análisis. Puedes usar Cloud Dataprep usando una IU basada en navegador, sin escribir código. Cloud Dataprep implementa y administra automáticamente los recursos necesarios para realizar las transformaciones, según la demanda.

Con Cloud Dataprep, puedes transformar datos de cualquier tamaño almacenados en formato CSV, JSON o tabla relacional. Cloud Dataprep usa Cloud Dataflow para escalar automáticamente y puede manejar conjuntos de datos a escala de terabyte. Como Cloud Dataprep está completamente integrado con Google Cloud Platform, puedes procesar datos sin importar dónde residan: en Cloud Storage, en BigQuery o en tu escritorio. Después de procesar los datos, puedes exportar los datos limpios directamente a BigQuery para un análisis posterior. Puedes gestionar el acceso de los usuarios y la seguridad de los datos con Cloud Identity y Access Management.

Aquí hay algunos casos de uso comunes para Cloud Dataprep:

- Aprendizaje automático: puedes limpiar los datos de entrenamiento para afinar los modelos ML.
- Análisis: puedes transformar datos brutos para que puedan ser utilizados en herramientas de almacenamiento de datos como BigQuery.

## **Analizando y consultando datos**

Una vez que los datos son ingeridos, almacenados y procesados, es necesario que terminen en un formato que permita acceder y consultar con facilidad.

BigQuery: data warehouse administrado

BigQuery es un data warehouse totalmente administrado con soporte para consultas SQL ad-hoc y esquemas complejos. Puedes usar BigQuery para analizar, comprender y organizar datos. Los clientes acostumbrados a usar un data warehouse tradicional para ejecutar consultas SQL estándar o herramientas de inteligencia empresarial y visualización apreciarán la interfaz de BigQuery.

BigQuery es un data warehouse OLAP de análisis altamente escalable, altamente distribuido y de bajo costo capaz de lograr una tasa de exploración de más de 1TB / seg. Es un servicio completamente administrado; los nodos computacionales se hilan para cada consulta ingresada en el sistema.

Para comenzar con BigQuery, crea un conjunto de datos dentro de tu proyecto, carga datos en una tabla y ejecuta una consulta. El proceso de carga de datos se puede simplificar utilizando la ingestión de transmisión desde Cloud Pub / Sub y Dataflow, cargando datos desde Cloud Storage, o utilizando la salida de un trabajo de procesamiento ejecutado en



Dataflow o Dataproc. BigQuery puede importar formatos de datos CSV, Avro y JSON e incluye soporte para elementos anidados y repetidos en JSON.

Todos los datos en BigQuery se acceden a través de un canal cifrado y se cifran en reposo. BigQuery está cubierto por los programas de cumplimiento de Google que incluyen SOC, PCI, ISO 27001 e HIPAA, por lo que se puede usar para manejar y consultar información confidencial. El acceso a los datos se controla a través de las ACL propiedad de los clientes.

BigQuery calcula los cargos de facturación a lo largo de dos dimensiones independientes: consultas y almacenamiento. Almacenar datos en BigQuery es comparable en costo con el almacenamiento de datos en Cloud Storage, lo que significa que no necesita elegir entre mantener los datos de registro en un depósito y en BigQuery. No existe un límite superior para la cantidad de datos que se pueden almacenar en BigQuery; además, si las tablas no se editan durante 90 días, el precio de almacenamiento para esa tabla disminuye en un 50%.

Un caso de uso típico para BigQuery es transmitir o cargar lotes de forma periódica datos de registro de servidores u otros sistemas que producen señales a gran velocidad, como los dispositivos IoT. La integración nativa está disponible con varios servicios de Google. Por ejemplo, el registro Stackdriver se puede configurar para entregar datos de registro directamente en BigQuery.

Al consultar datos en BigQuery, tienes la opción de 2 modelos de precios: on-demand o flat-rate. Con los precios on-demand, los cargos por consulta tienen un precio de acuerdo con Terabytes procesados. Al usar precios flat-rate, BigQuery te brinda capacidad de consulta consistente con un modelo de costes más simple.

Como un servicio completamente administrado, BigQuery automatiza tareas tales como ventanas de mantenimiento de infraestructura y aspiración de datos. Para mejorar el diseño de tus consultas, puedes examinar la explicación del plan de consulta de cualquier consulta determinada. Los datos se almacenan en un formato de columna, que está optimizado para agregaciones a gran escala y procesamiento de datos. Además, BigQuery tiene soporte integrado para la partición de datos de series de tiempo. Desde una perspectiva de diseño, esto significa que puedes diseñar tu actividad de carga para usar una marca de tiempo y luego buscar consultas en una partición de fecha particular. Debido a que los cargos de consulta de BigQuery se basan en la cantidad de datos escaneados, la partición adecuada de los datos puede mejorar en gran medida la eficiencia de la consulta y reducir los costes.

La ejecución de consultas en BigQuery se puede hacer utilizando SQL estándar, que es compatible con SQL 2011 y tiene extensiones para permitir la consulta de datos anidados y repetidos. Existe un amplio conjunto de funciones incorporadas y operadores nativamente disponibles dentro de BigQuery, y soporte para funciones definidas por el usuario (UDF).

Puedes aprovechar BigQuery de varias formas; considera los siguientes casos de uso:

- **Análisis del usuario:** ingiere grandes cantidades de actividad generada por el usuario (adtech, clickstream, telemetría del juego) y determina el comportamiento y las características del usuario.
- **Métricas operativas y de dispositivos:** recopila información de transmisión de sistemas de TI, dispositivos de IoT, etc. y analiza datos de tendencias y variaciones.
- **Inteligencia empresarial:** almacena las métricas de negocio como un depósito de datos e impulse una herramienta de BI o una oferta de socio, como Tableau, QlikView o Looker.

### **Comprender los datos con machine learning**

El machine learning se ha convertido en un componente crítico de la fase de análisis del ciclo de vida de los datos. Se puede utilizar para aumentar los resultados procesados, sugerir optimizaciones de recopilación de datos y predecir los resultados dentro de los conjuntos de datos.

Considera los siguientes casos de uso:

- **Recomendaciones de producto:** puedes crear un modelo que recomiende productos para clientes en función de compras previas y navegación del sitio.
- **Predicción:** utiliza el aprendizaje automático para predecir el rendimiento de sistemas complejos, como los mercados financieros.
- **Asistentes automáticos:** genera asistentes automáticos que entiendan y respondan a las preguntas de los usuarios.
- **Análisis de opinión:** determina el sentimiento subyacente de los comentarios de los usuarios sobre las reseñas de productos y las noticias.

Hay varias opciones para aprovechar el machine learning dentro de Cloud Platform.

- **API de machine learning para tareas específicas:** Google Cloud Platform brinda servicios de machine learning administrados y llave en mano con modelos preparados para la visión, el habla, el lenguaje natural y la traducción de textos. Estas API se crean a partir de las mismas tecnologías que potencian aplicaciones como Google Photos, la aplicación móvil de Google, Google Translate y las respuestas inteligentes de Inbox.

- machine learning personalizado: Cloud Machine Learning Engine es un servicio alojado y administrado que ejecuta modelos personalizados a escala. Además, Cloud Dataproc también puede ejecutar modelos de machine learning construidos con Mahout o Spark MLlib.

### **API de Cloud Vision**

Puedes utilizar la API de Cloud Vision para analizar y comprender el contenido de una imagen utilizando redes neuronales previamente capacitadas. Con la API de Cloud Vision puede clasificar imágenes, detectar objetos y caras individuales, y reconocer palabras impresas. Además, puedes usar la API de Cloud Vision para detectar contenido inapropiado y analizar los atributos faciales emocionales de las personas.

La API de Cloud Vision es accesible a través de un API REST. Puedes enviar imágenes directamente al servicio o subirlas a Cloud Storage e incluir un enlace a la imagen en la solicitud. Las solicitudes pueden incluir una sola imagen, o múltiples imágenes pueden ser anotadas en un solo lote. Dentro de una solicitud, se pueden seleccionar anotaciones de características específicas para la detección de cada imagen adjunta. La detección de características incluye etiquetas, texto, caras, puntos de referencia, logotipos, búsqueda segura y propiedades de imagen (como colores dominantes). La respuesta contendrá metadatos sobre cada anotación de tipo de característica seleccionada para cada una de las proporcionadas en la solicitud original.

Puede integrar fácilmente la API de Cloud Vision en aplicaciones personalizadas que se ejecutan en App Engine, Kubernetes Engine, Compute Engine y plataformas móviles como Android e iOS. También se puede acceder desde los servicios de Google Cloud Platform como Cloud Dataflow, Cloud Dataproc y Cloud Datalab o Notebooks de Jupyter.

### **Cloud Speech API**

La API de Cloud Speech admite la capacidad de analizar audio y convertirlo en texto. La API reconoce más de 80 idiomas y variantes y está potenciada por algoritmos de redes neuronales de aprendizaje profundo que evolucionan y mejoran constantemente.

Puede usar la API de Cloud Speech para diferentes tipos de cargas de trabajo:

- Paso a texto en tiempo real: Cloud Speech API puede aceptar la entrada de audio en tiempo real y comenzar a devolver los resultados de reconocimiento parcial a medida que estén disponibles. Esta capacidad es útil para integrar el dictado en tiempo real o habilitar el comando y control a través de la voz dentro de las aplicaciones. Cloud Speech API es compatible con gRPC, un marco RPC de alto rendimiento, de código abierto y general, para la transmisión de análisis de audio y voz para aplicaciones personalizadas que se ejecutan en App Engine, Kubernetes Engine, Compute Engine y plataformas móviles como Android e iOS. .

- **Análisis por lotes:** para procesar grandes cantidades de archivos de audio, puedes llamar a la API de Cloud Speech utilizando los puntos finales REST y gRPC. Se admiten las capacidades sincrónicas y asíncronas de voz a texto. También se puede acceder a la API REST desde servicios Google Cloud Platform como Cloud Dataflow, Cloud Dataproc y Cloud Datalab Notebooks de Jupyter.

### **API de lenguaje natural en la nube**

Cloud Natural Language API proporciona la capacidad de analizar y revelar la estructura y el significado del texto. La API se puede usar para extraer información sobre personas, lugares, eventos, el sentimiento del texto de entrada y más. El análisis resultante se puede utilizar para filtrar contenido inapropiado, clasificar el contenido por temas o crear relaciones a partir de las entidades extraídas que se encuentran en el texto de entrada.

Puede combinar la API de lenguaje natural con las capacidades de OCR de la API de Cloud Vision o las funciones de voz a texto de la API de Cloud Speech para crear aplicaciones o servicios.

La API de lenguaje natural está disponible a través de un API REST. Puede enviar texto directamente al servicio o cargar archivos de texto a Cloud Storage y vincularlo con el texto de su solicitud. Puede integrar fácilmente la API en aplicaciones personalizadas que se ejecutan en App Engine, Kubernetes Engine, Compute Engine y plataformas móviles como Android e iOS. También se puede acceder desde otros servicios de Google Cloud Platform como Cloud Dataflow, Cloud Dataproc o Cloud Datalab Notebooks de Jupyter.

### **API de traducción en la nube**

Puedes usar la API de Cloud Translation para traducir más de 90 idiomas diferentes. Si el idioma de entrada es desconocido, la API de Cloud Translation detecta automáticamente el idioma con gran precisión.

La API de Cloud Translation puede proporcionar traducción en tiempo real para aplicaciones web y móviles, y admite solicitudes por lotes para cargas de trabajo analíticas.

La API de Cloud Translation está disponible a través de API REST. Puede integrar fácilmente la API en aplicaciones personalizadas que se ejecutan en App Engine, Kubernetes Engine, Compute Engine y plataformas móviles como Android e iOS. También se puede acceder desde los servicios de Google Cloud Platform como Cloud Dataflow, Cloud Dataproc o Cloud Datalab Notebooks de Jupyter.

### **Cloud Video Intelligence API: búsqueda de video**

El contenido de video ha sido tradicionalmente opaco y no se ha prestado fácilmente al análisis. Pero con la API Cloud Video Intelligence, una API REST fácil de usar, ahora puedes buscar, descubrir y extraer metadatos de los videos. Cloud Video Intelligence puede detectar entidades (nombres) en el contenido de video, como "perro", "flor" o "automóvil". También puede buscar entidades dentro de escenas específicas en el contenido del video.

Los usuarios pueden anotar videos con metadatos a nivel de marco y video. (El servicio puede extraer datos con una granularidad máxima de 1 fotograma por segundo). La API admite formatos de video comunes, incluidos MOV, MPEG4, MP4 y AVI.

Hacer una solicitud para anotar un video es sencillo: crea un archivo de solicitud JSON con la ubicación del video y el tipo o tipo de anotación que desea realizar, y luego envía la solicitud al API.

Estos son algunos casos de uso comunes para Cloud Video Intelligence:

- Obten información de los videos: extrae información de los videos sin tener que usar el machine learning o implementar algoritmos de visión por computadora.
- Búsqueda por catálogo de video: busca a través de un catálogo de videos para identificar la presencia y la marca de tiempo de las entidades de interés.

### **Cloud Machine Learning: plataforma de machine learning gestionado**

Cloud Machine Learning Engine es una plataforma administrada que puede usar para ejecutar modelos personalizados de machine learning a escala. Puedes crear modelos utilizando el marco TensorFlow, un marco de código abierto para inteligencia artificial, y luego usar Cloud Machine Learning para administrar el preprocesamiento, el entrenamiento y la predicción.

Cloud ML Engine está integrado con Cloud Dataflow para el procesamiento previo de datos, que puede acceder a los datos almacenados en Cloud Storage y BigQuery. También funciona con Cloud Load Balancer para servir predicciones en línea a escala.

Puedes desarrollar y probar modelos TensorFlow por completo dentro de Google Cloud Platform utilizando los notebooks de Jupyter, y luego usar Cloud Machine Learning para entrenamientos a gran escala y cargas de trabajo de predicción.

Los modelos creados para Cloud Machine Learning son completamente portátiles. Al aprovechar el marco TensorFlow, puede construir y probar modelos localmente y luego implementarlos en múltiples máquinas para entrenamiento distribuido y predicción. Finalmente, puede cargar los modelos capacitados en Cloud Machine Learning y ejecutarlos en varias instancias distribuidas de máquinas virtuales.

El flujo de trabajo de Cloud Machine Learning consiste en las siguientes fases:

- **Preprocesamiento:** Cloud Machine Learning convierte las características de los conjuntos de datos de entrada en un formato compatible, y también puede normalizar y transformar los datos para permitir un aprendizaje más eficiente. Durante el preprocesamiento, los datos de entrenamiento, evaluación y prueba se almacenan en Cloud Storage. Esto también hace que los datos sean accesibles a Cloud Dataflow durante esta fase para cualquier preprocesamiento requerido adicional.
- **Creación de gráficos:** Cloud Machine Learning convierte el modelo TensorFlow suministrado en un modelo de Cloud Machine Learning con operaciones de capacitación, evaluación y predicción.
- **Training-Cloud Machine Learning** continuamente itera y evalúa el modelo de acuerdo a los parámetros presentados.
- **Predicción:** Cloud Machine Learning utiliza el modelo para realizar cálculos. Las predicciones se pueden calcular en lotes o bajo demanda, como un servicio de predicción en línea. Las predicciones por lotes están diseñadas para ejecutarse de forma asincrónica con grandes conjuntos de datos, utilizando servicios como Cloud Dataflow para orquestar el análisis. Las predicciones a pedido a menudo se usan con aplicaciones personalizadas que se ejecutan en App Engine, Kubernetes Engine o Compute Engine.

### **Machine Learning de propósito general**

Además de la API y la plataforma de aprendizaje automático construidas por Google, puedes implementar otras herramientas de Machine Learning de gran escala en Google Cloud Platform. Mahout y MLlib son dos proyectos dentro de los ecosistemas Hadoop y Spark, que proporcionan una gama de algoritmos de Machine Learning de propósito general. Ambos paquetes ofrecen algoritmos de Machine Learning para clustering, clasificación, filtrado colaborativo y más.

Puedes usar Cloud Dataproc para implementar clústeres administrados de Hadoop y Spark, y arrancar esos clústeres con software adicional. Esto significa que puedes ejecutar cargas de trabajo de Machine Learning compiladas con Mahout o MLlib en Google Cloud Platform, y poder escalar los clústeres utilizando máquinas virtuales normales o prioritarias.

### **Explorar y visualizar**

El último paso en el ciclo de vida de los datos es la exploración y visualización de datos en profundidad para comprender mejor los resultados del procesamiento y análisis.

Los conocimientos adquiridos durante la exploración se pueden utilizar para impulsar mejoras en la velocidad o el volumen de la ingestión de datos, el uso de diferentes medios de almacenamiento para acelerar el análisis y mejoras en el procesamiento de las tuberías. Explorar y comprender completamente estos conjuntos de datos a menudo implica los servicios de científicos de datos y analistas de negocios, personas capacitadas en probabilidad, estadísticas y comprensión del valor comercial.

### **Explorando resultados de ciencia de datos**

La ciencia de datos es el proceso de obtener valor de los activos de datos en bruto. Para hacerlo, un científico de datos puede combinar conjuntos de datos dispares, algunos públicos, algunos privados, y realizar una gama de técnicas de agregación y análisis. A diferencia del almacenamiento de datos, los tipos de análisis y la estructura de los datos varían ampliamente y no están predeterminados. Las técnicas específicas incluyen métodos estadísticos, como agrupamiento, Bayesiano, máxima verosimilitud y regresión, así como Machine Learning, como árboles de decisión y redes neuronales.

### **Cloud Datalab, Notebooks de Jupyter: información de datos interactivos**

Cloud Datalab Notebooks de Jupyter es una herramienta interactiva basada en la web que puede usar para explorar, analizar y visualizar datos. Está construido sobre los notebooks de Jupyter, que anteriormente se conocía como IPython. Con Cloud DataLab Notebooks de Jupyter, puede, con un solo clic, iniciar un notebook interactivo basado en la web donde los usuarios pueden escribir y ejecutar programas de Python para procesar y visualizar datos. Los notebooks mantienen su estado y se pueden compartir entre los científicos de datos, así como publicados en sitios como GitHub, Bitbucket y Dropbox.

Desde el primer momento, incluye soporte para muchos kits de herramientas populares de ciencia de datos, incluyendo pandas, numpy y scikit-learn, y paquetes de visualización comunes, como matplotlib. También incluye soporte para Tensorflow y Cloud Dataflow. Usando estas bibliotecas y servicios en la nube, un científico de datos puede cargar y limpiar datos, construir y verificar modelos, y luego visualizar los resultados usando matplotlib. Esto funciona tanto para datos que se ajustan a una sola máquina como para datos que requieren un clúster para almacenar. Se pueden cargar módulos adicionales de Python usando los comandos de instalación de! Pip.

### **Ecosistema para el científico de datos**

Con las instancias de Compute Engine de alto rendimiento, puede implementar muchos tipos de herramientas de ciencia de datos y usarlas para ejecutar análisis a gran escala en Google Cloud Platform.

El lenguaje de programación R es comúnmente utilizado por los estadísticos. Si desea utilizar R para la exploración de datos, puede implementar RStudio Server o Microsoft R Server en una instancia de Compute Engine. RStudio Server proporciona un entorno de tiempo de ejecución interactivo para procesar y manipular datos, crear modelos sofisticados y visualizar resultados. Microsoft R Server es un complemento a gran escala y de alto rendimiento para los clientes de escritorio R para ejecutar cargas de trabajo analíticas.

Cloud Datalab está basado en Jupyter y actualmente es compatible con Python. Si deseas realizar tu exploración de datos en otros idiomas, como R, Julia, Scala y Java, puedes implementar Jupyter de código abierto o JupyterHub en instancias de Compute Engine.

Apache Zeppelin es otra herramienta popular de ciencia de datos, basada en notebooks y basada en la web. De forma similar a Jupyter, Zeppelin proporciona soporte para sistemas de back-end de procesamiento de datos y lenguaje adicionales como Spark, Hive, R y Python.

Tanto Jupyter como Zeppelin se pueden implementar utilizando acciones de inicialización de Cloud Dataproc preconstruidas para iniciar rápidamente paquetes de software Hadoop y Spark-ecosystem comunes.

## **Visualizando resultados de BI**

Durante la fase de análisis, puede resultar útil generar visualizaciones de datos, cuadros de mando e informes complejos para explicar los resultados del procesamiento de datos a un público más amplio. Para hacerlo más fácil, Google Cloud Platform se integra con una serie de herramientas de creación de informes y cuadros de mando.

Google Data Studio proporciona un generador de informes de arrastrar y soltar que puede usar para visualizar datos en informes y paneles que luego pueden compartirse con otras personas. Los cuadros y gráficos en los informes están respaldados por datos en tiempo real, que se pueden compartir y actualizar fácilmente. Los informes pueden contener controles interactivos que permiten a los colaboradores ajustar las dimensiones utilizadas para generar visualizaciones.

Con Data Studio puedes crear informes y cuadros de mando a partir de archivos de datos existentes, Hojas de cálculo de Google, Google Cloud SQL y BigQuery. Al combinar Data Studio con BigQuery, puedes aprovechar toda la capacidad informática y de almacenamiento de BigQuery sin tener que importar datos manualmente a Data Studio o crear integraciones personalizadas.



Si prefieres visualizar datos en una hoja de cálculo, puedes usar Hojas de cálculo de Google, que se integra directamente con BigQuery. Con Google Apps Script, puedes insertar consultas y datos de BigQuery directamente en Hojas de cálculo de Google. También puedes exportar resultados de consultas de BigQuery en archivos CSV y abrirlos en Hojas de cálculo de Google u otras hojas de cálculo. Esto es útil para crear conjuntos de datos más pequeños para compartir o analizar. También puedes hacer lo contrario, usar BigQuery para consultar entre conjuntos de datos distribuidos almacenados en Hojas de cálculo de Google o archivos almacenados en Google Drive.

BigQuery también es compatible con una amplia gama de herramientas e integraciones de inteligencia empresarial de terceros, que van desde SaaS hasta aplicaciones de escritorio.

## Orquestación

La incorporación de todos los elementos del ciclo de vida de los datos en un conjunto de operaciones conectadas y cohesivas requiere alguna forma de orquestación. Las capas de orquestación generalmente se utilizan para coordinar las tareas de inicio, detener tareas, copiar archivos y proporcionar un panel para supervisar los trabajos de procesamiento de datos. Por ejemplo, un flujo de trabajo podría incluir la copia de archivos en Cloud Storage, el inicio de un trabajo de procesamiento de Cloud Dataproc y luego el envío de notificaciones cuando los resultados del procesamiento se almacenan en BigQuery.

Los flujos de trabajo de orquestación pueden variar de simples a complejos, dependiendo de las tareas de procesamiento, y a menudo usan un mecanismo de programación centralizado para ejecutar flujos de trabajo automáticamente. Hay varias herramientas de orquestación de código abierto que admiten Google Cloud Platform, como Luigi y Airflow. Para las aplicaciones de orquestación personalizadas, puede crear una aplicación de App Engine que use la funcionalidad integrada de tareas programadas de App Engine para crear y ejecutar flujos de trabajo.

## Resumen

