GUÍA SOBRE EL APRENDIZAJE AUTOMÁTICO Y EL ANÁLISIS DE DATOS PARA DIRECTORES DE TI

Google Cloud

CONTENIDO

INTRODUCCIÓN

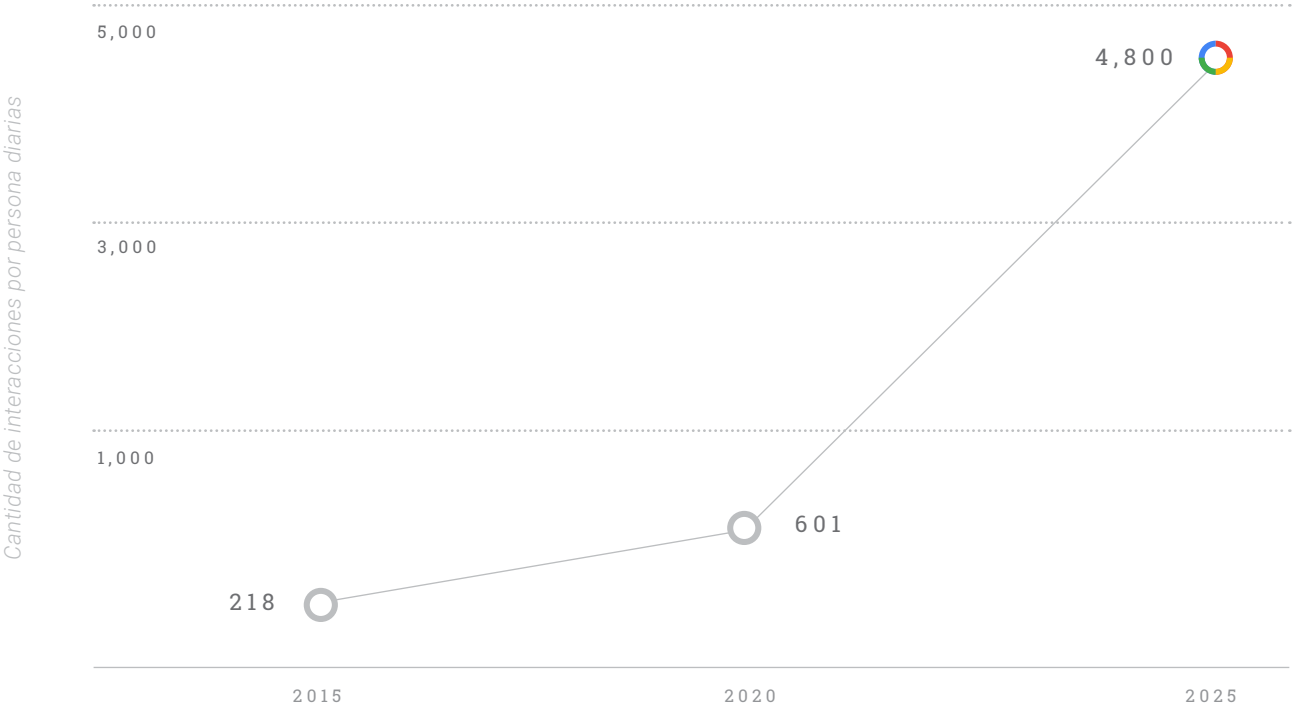
El uso de los datos para tomar decisiones comerciales ya no es ninguna novedad. Antes, la "toma de decisiones basada en datos" podía referirse al hecho de notar una correlación entre una campaña de anuncios impresos y los relatos anecdóticos de una cantidad de ventas mayor que la habitual. Las empresas usaban todos los datos que podían obtener, en el momento en que podían obtenerlos.

En la actualidad, los datos se encuentran en todas partes. Se transmiten a velocidades impresionantes desde dispositivos conectados, se encuentran en una variedad de formatos y provienen de miles de millones de usuarios. Los macrodatos suelen considerarse una oportunidad, pero solo las empresas con la estructura necesaria para manejar su volumen y diversidad la pueden aprovechar. Para otras empresas, la cantidad enorme de datos puede representar un riesgo: que no se aprovechen las posibles estadísticas, no se satisfagan las necesidades de los clientes y se sigan tomando decisiones sin información.

Hay dos factores que hacen que el panorama actual sea diferente al de los avances del pasado. El primero es un crecimiento *exponencial del volumen y la diversidad de los datos que provienen de* miles de millones de usuarios y dispositivos. El segundo es la *demanda de acceso inmediato a estadísticas y datos de alta calidad.* Estos factores generaron un sentido de urgencia en las empresas sobre la manera en la que administran los datos. Además, el costo y el rendimiento de muchas capacidades de la nube alcanzaron un punto de inflexión, que permitió el acceso de todas las empresas al aprendizaje automático (ML) y la inteligencia artificial (IA).

A pesar del reconocimiento generalizado del valor de los datos, pocas empresas han implementado las estrategias de datos modernas.[1] Esta guía, basada en las investigaciones originales y las contribuciones de Google en la nube, se diseñó para ayudar a los líderes empresariales y de TI a implementar estrategias para la administración de datos modernas y basadas en la nube. En cada sección, destacamos las tecnologías que permiten que las empresas conviertan un panorama de datos complejo y vasto en estadísticas empresariales útiles.

INTERACCIONES POR PERSONA CONECTADA DIARIAS



En 2025, la persona conectada promedio interactuará con dispositivos conectados casi 4,800 veces por día, lo que equivale a una interacción cada 18 segundos.[2]

NUESTROS INICIOS

La *guía sobre el aprendizaje automático y el análisis de datos* de Google Cloud se basa en los veinte años en que Google ha abordado algunos de los problemas de datos más complicados de la industria. A lo largo del camino, contribuimos con investigaciones originales que ayudaron a dar forma al panorama de los macrodatos, desde los dos documentos de investigación de finales de 2003 y 2004, que generaron el movimiento Hadoop, hasta el informe de Dremel, que estableció la base del almacén de datos en nube que se describe en esta guía.

Diseñamos, compilamos e implementamos Spanner, el primer sistema destinado a distribuir datos a escala global y respaldar las transacciones distribuidas y coherentes de manera externa. En 2017, se puso a disposición general de nuestros clientes.[3] Recientemente, Google Brain ayudó a fomentar el interés renovado de la industria en la IA, lo que generó el lanzamiento de nuestro proyecto TensorFlow en código abierto.[4] Con esta guía, buscamos compartirles nuestra experiencia a los líderes que desean aprovechar el aprendizaje automático y la IA para sus organizaciones.

CAPÍTULO 1

EL NUEVO PANORAMA DE DATOS

01

01

EL NUEVO PANORAMA DE DATOS

La administración de datos sería más fácil si el crecimiento se limitara solo a algunos recursos o si los datos fueran uniformes. El desafío está en la diversidad de las fuentes y los formatos. Esa diversidad incluye el volumen creciente de datos no estructurados: correos electrónicos, registros de sistemas, páginas web, transcripciones de clientes, documentos, diapositivas, conversaciones informales y el notorio aumento de rich media, como los videos y las imágenes en HD. Desde cualquier dispositivo conectado a Internet, se puede acceder a grandes volúmenes de información, lo que genera nuevas expectativas sobre la inmediatez y la disponibilidad de los datos.

Las aplicaciones para consumidores, como la búsqueda, la mensajería, el comercio electrónico, las redes sociales y los videos en línea fueron los primeros en tener este problema. Se tuvieron que crear sistemas nuevos para manejar el tráfico a la escala de toda la Web y obtener estadísticas al instante. Ahora, estos avances, que son cada vez más importantes, están disponibles para todas las empresas y permiten tanto ayudar a los fabricantes a administrar cadenas de suministros de manera más eficaz como a aumentar la precisión de los diagnósticos médicos.

Los equipos de TI se mantienen en el medio. Deben descubrir una manera de proveer una *vista en tiempo real de la empresa* y, al mismo tiempo, administrar *un panorama de datos más grande y complejo*. Como ocurre con muchas iniciativas de software, reducir la complejidad es un factor decisivo para lograr el éxito.

En esta guía, se analiza la manera en que los servicios de nube administrados les permiten a las empresas nuevas y a las ya establecidas abordar los desafíos de datos de la actualidad. Presenta una ruta que comienza con la recopilación de datos empresariales sin procesar para almacenarlos en la nube. A medida que surgen los interrogantes de las empresas, las herramientas basadas en la nube pueden preparar

y estructurar a pedido los datos sin procesar. Luego, los datos preparados se integran en un almacén de datos en la nube, al que se puede acceder de inmediato para su análisis. Este tesoro de datos sirve como una "base" para que las organizaciones puedan recopilar, preparar y analizar datos de todo tipo y de cualquier fuente. La naturaleza completamente administrada de los servicios de nube permite optimizar todo este proceso, incluida la asistencia para el análisis en tiempo real, sin la necesidad de que TI esté al tanto de la infraestructura subyacente. Con esta base como punto de partida, la conclusión de esta guía muestra cómo las organizaciones pueden usar este ciclo de recopilación y preparación de datos para habilitar el aprendizaje automático y la IA.

SIN SERVIDORES: LA RUTA A LA PRODUCTIVIDAD DE TI

Las arquitecturas modernas sin servidores son el resultado de una serie de iniciativas realizadas para *reducir la cantidad de responsabilidades* que tienen los programadores y equipos de TI. Fundamentalmente, el objetivo de la computación sin servidores es quitar el trabajo comercializado (administración de clústeres de servidores, fragmentación de bases de datos, balanceo de cargas, planificación de capacidades, aseguramiento de disponibilidad, entre otros) para que los equipos de TI puedan enfocarse en lo que es importante para la empresa. La computación sin servidores genera una gran diferencia entre la *TI comercializada* (el trabajo de mantenimiento rutinario que prácticamente es el mismo en cada empresa) y el *trabajo diferenciado* que convierte a TI en un proveedor directo de valor comercial.

RESUMEN DEL CAPÍTULO 1

1 **Las empresas enfrentan 3 desafíos nuevos:**

- el *volumen* de los datos que se crean
- la *diversidad* de las fuentes y los formatos de los datos
- la *velocidad* con la que los consumidores y las partes interesadas internas esperan las estadísticas en la actualidad

2 **La computación en la nube** les permite a las empresas superar desafíos mediante la *administración de datos a escala y con velocidad,* sin necesidad de preocuparse por la infraestructura.

3 **Específicamente, las empresas pueden comenzar a modernizar sus estrategias de datos** mediante el enfoque en el *almacenamiento en la nube y el almacén de datos* como primer paso para el establecimiento de una base para el aprendizaje automático y la IA.

CASO DE CLIENTE

FIS

FIS usa los servicios de nube completamente administrados para analizar los factores disruptivos y los eventos del mercado

FIS desarrolló una herramienta de reconstrucción de mercado que podría ayudar a determinar la posible causa de los eventos que alteran el mercado de valores, como el "colapso relámpago" de 2010. El sistema de FIS no solo puede almacenar miles de millones de transacciones, sino que también les permite a los ejecutivos encargados del riesgo y el cumplimiento llevar a cabo la vigilancia y las consultas a pedido, incluida la reconstrucción del ercado.

Según FIS, el sistema puede procesar y vincular hasta 15 terabytes de datos diarios en cuatro horas y puede almacenarlos por seis años, en conformidad con lo exigido por la ley. "La cifra total es de unos 30 petabytes de datos", señala Neil Palmer, director de Tecnología de asuntos de tecnología avanzada en FIS. "No hay mucha oferta para esa escala y, definitivamente, no hay en los servicios financieros. Es una gran iniciativa".

El equipo de Palmer necesitaba una plataforma con una gran potencia de procesamiento, pero también quería evitar el costo y mantenimiento asociados con el desarrollo y la operación de un sistema in situ. "La escalabilidad es un gran beneficio que se obtiene de Google Cloud Platform", afirma Palmer. "Una solución tradicional de TI con hardware fijo en este escenario podría dar como resultado que equipos de miles de millones de dólares estén inactivos durante muchos días hábiles".

MÁS INFORMACIÓN

EMPRESA

FIS

INDUSTRIA

Servicios financieros

ACERCA DE

FIS es un líder global en tecnología de servicios financieros. Se centra en las ventas minoristas y la banca institucional, los pagos, la administración de recursos y bienes, la gestión del riesgo y el cumplimiento, la consultoría y las soluciones de subcontratación.

CAPÍTULO 2

ALMACENAMIENTO EN LA NUBE Y ALMACÉN DE DATOS

02

ALMACENAMIENTO EN LA NUBE Y ALMACÉN DE DATOS

02

Uno de los primeros pasos que las organizaciones pueden tomar para modernizarse es almacenar en la nube los datos sin procesar de los procesos empresariales clave. Al hacerlo, pueden aprovechar las capacidades de análisis en la nube.

Los sistemas de datos aislados que están repartidos en las empresas siguen siendo un problema para los equipos empresariales y de TI, ya que se crean nuevos sistemas aislados (por motivos organizacionales o técnicos) diariamente.[5] *Harvard Business Review* publicó material sobre la necesidad de una única fuente confiable de datos y un medio claro mediante el cual las distintas líneas empresariales puedan ver los datos.[6]

El almacenamiento en la nube y el almacén de datos les permiten a las empresas mantener un repositorio único y central, y habilitar funciones empresariales diferentes para analizar los datos en maneras que satisfagan sus necesidades únicas, con una mayor velocidad y flexibilidad que antes. Juntas, estas capacidades permiten crear una vista general de la empresa en todos los sistemas aislados.

Recopila datos sin procesar para realizar análisis futuros

IDC estima que se analiza menos del 1% de todos los archivos.[7] El otro 99%, según el momento de las necesidades empresariales, contiene material útil para obtener estadísticas que permitan dirigir la toma de decisiones. Las organizaciones no pueden predecir los interrogantes empresariales que surgirán, es por eso que necesitan almacenar grandes cantidades de datos con rapidez y de forma económica y flexible. En especial con los archivos no estructurados, que son la mayoría de los datos que se generan.[8]

Con la nube, las empresas pueden almacenar enormes cantidades de archivos a un costo bajo (menos de un centavo de dólar por gigabyte en el momento de la redacción de este documento).[9] Los datos que se necesiten actualmente se pueden mantener "preparados" (disponibles de manera global para funcionar en aplicaciones o ejecutar análisis), al mismo tiempo en que se usa un almacenamiento "en frío" más barato de los datos con valores aún sin procesar. El almacenamiento en línea más poderoso es el que permite que incluso los datos inactivos archivados se recuperen con rapidez y con una latencia extremadamente baja.

El almacenamiento en la nube no solo permite ahorrar dinero, sino que también sirve como base para la realización de análisis importantes. Las empresas pueden recopilar archivos estructurados y no estructurados sin complicaciones y en sus formatos nativos. Debido a que el almacenamiento se separa intencionalmente del procesamiento y el análisis, los equipos pueden posponer el estructuramiento de los datos sin procesar destinados a los análisis hasta que surjan las interrogantes empresariales. Fundamentalmente, los datos sin procesar que provienen de la misma base pueden reestructurarse con facilidad para responder las preguntas nuevas que surjan. Lo que diferencia al almacenamiento en la nube es el nivel de eficacia en el que ocurren los pasos de reutilización y recopilación de datos. Para lograr que una organización se beneficie de los análisis, los equipos deben asegurarse de que se recopilen y centralicen los datos sin procesar de sus procesos empresariales.

Esta flexibilidad está acelerando la adopción de la nube como repositorio de los datos no estructurados de las organizaciones: cerca de la mitad de las organizaciones en Asia-Pacífico, Estados Unidos y Europa anticipan alzas de al menos el 5% en su almacenamiento de datos no estructurados en la nube durante el próximo año. Varias informaron un aumento mayor al 10%.[10]

EL INTERNET DE LAS COSAS

Según una encuesta realizada a más de 500 líderes de TI globales por *MIT Sloan Management Review* en nombre de Google Cloud, la adopción de la nube sigue acelerándose. Se espera que la mayoría (65%) de las aplicaciones, datos o infraestructura se base en la nube en 2019.

El Internet de las cosas (IoT) es un factor importante de este traslado a la nube, ya que actualmente un 91% de los encuestados con iniciativas de IoT implementan (59%) o están planificando implementar (32%) los datos de los dispositivos conectados a IoT en la nube. Los encuestados mencionaron que la capacidad de integrar nuevas herramientas y plataformas (33%), la implementación e iteración más rápidas (31%), la flexibilidad incrementada en los procesos empresariales y las opciones de los proveedores (29%), y una mayor seguridad (28%) son las principales razones para implementar los datos de IoT en la nube.

Para usar los datos de IoT de manera significativa, las empresas deben ser capaces de comprenderlos en contexto. Un almacén de datos en la nube que permite entradas por lote y de transmisión, junto con una plataforma de análisis poderosa, ayuda a garantizar que tus datos de IoT puedan proporcionar estadísticas en tiempo real.

Administra los datos en los sistemas aislados

Con la capacidad de capturar datos de cualquier tipo en forma económica, las organizaciones pueden centrar sus esfuerzos en el desarrollo de una mirada disciplinada de sus procesos empresariales más importantes. Así como el almacenamiento en la nube centraliza datos en su formato nativo sin procesar, un almacén de datos en la nube les permite a las empresas recopilar datos de sistemas aislados diversos para realizar análisis, como lo haría un almacén de datos tradicional. Con la nube, las empresas pueden administrar grandes cantidades de datos con inversiones de capital mínimas, escalar de manera prácticamente indefinida y solo pagar por lo que usan. Los servicios de nube administrados van un paso más adelante, ya que permiten que TI no se deba encargar de la infraestructura subyacente. Las empresas deben considerar los interrogantes empresariales que deben responderse y los datos necesarios para dar esas respuestas.

Estos son algunos ejemplos:

- ¿Cuáles son los objetivos comerciales principales de mis datos? ¿Acaso es conocer cómo los usuarios interactúan con mis sistemas, identifican tendencias, aumentan las ventas, desarrollan la lealtad del consumidor o es algo más?
- ¿De dónde provendrán mis datos más importantes (transacciones, registros de servidores, servicios de nube, dispositivos/IoT, medios sociales)? ¿Ya se importaron al almacenamiento en la nube?
- ¿Qué tan rápido mi sistema debe incorporar datos nuevos en los informes y las visualizaciones?
- ¿Existe en la organización una cultura que fomente la toma de decisiones basada en datos (no solo entre los analistas de TI y científicos de datos)? ¿Quién debería tener acceso a la plataforma de análisis?

Una vez que la empresa determine sus objetivos comerciales, debe identificar las fuentes de datos de entrada en los sistemas aislados para importarlos al almacén de datos en la nube a fin de analizarlos. Esta es una lista de fuentes de entrada típicas:

Almacenamiento en la nube
Los datos almacenados en la nube se pueden importar a un almacén de datos en la nube para la realización de análisis.[12] En esta etapa, se puede formalizar un esquema según los interrogantes empresariales que se deban responder, lo que les provee una estructura a los datos sin procesar para el análisis.

Bases de datos transaccionales y de análisis
Los datos almacenados en bases de datos transaccionales y de análisis se pueden cargar por lotes o transmitir fila por fila directamente a un almacén de datos en la nube.

Los datos almacenados en los servicios de nube
Los datos almacenados con proveedores de SaaS populares se pueden importar a un almacén de datos en la nube, en muchos casos de forma automática.

Transmisión de datos
Los datos que provienen de la Web, los dispositivos móviles y las aplicaciones de IoT pueden eludir el almacenamiento en la nube y transmitirse directamente a un almacén de datos en la nube (consulta el capítulo 3: Integración de datos en tiempo real).

Control de los datos

El crecimiento exponencial en el volumen global de los datos no es el único obstáculo que enfrentan las empresas. Según Forrester, los requisitos cambiantes para el análisis y la generación de informes, y la desalineación existente entre la empresa y TI se encuentran entre los principales desafíos que obstaculizan las iniciativas en inteligencia empresarial de las organizaciones.[13] Además, la brecha en el talento para la ciencia de datos, que se ha documentado extensamente, (**consulta "Surgimiento del ciudadano científico de datos"**) exige que las empresas consideren enfoques nuevos para desarrollar conocimientos analíticos.

Según el acceso que corresponda a cada función, cualquier persona o programador de aplicaciones puede consultar datos almacenados en un almacén de datos de nube, generar informes o acceder a visualizaciones. Los almacenes de datos en la nube admiten la administración del acceso individualizado de la que se debe tener conocimientos. Los controles de acceso personalizados y la auditabilidad completa permiten democratizar la ciencia de datos, al mismo tiempo en que se mantienen las medidas de seguridad. De hecho, más de la mitad de las empresas en Asia-Pacífico, Estados unidos y Europa informan que están implementando, implementaron o están expandiendo el uso de las herramientas de inteligencia artificial empresarial de autoservicio en la organización.[14]

SURGIMIENTO DEL CIUDADANO CIENTÍFICO DE DATOS

Antes, obtener conclusiones precisas en términos estadísticos era responsabilidad exclusiva de los científicos de datos profesionales. Sin embargo, McKinsey afirma que, en 2018, "Estados Unidos podría sufrir un déficit de 140,000 a 190,000 personas con habilidades de análisis sólidas y 1.5 millones de administradores y analistas con conocimientos sobre cómo usar el análisis de los macrodatos para tomar decisiones eficaces".[15]

A medida que se intensifica la competencia, la mayoría de las empresas necesitarán una estrategia de talento diversificada. De acuerdo con InformationWeek *los ciudadanos científicos de datos* son personas que aprovechan el análisis de datos, pero cuyas funciones principales no se relacionan con las estadísticas ni el análisis. Estos profesionales pueden ser un complemento importante para los científicos de datos internos, sobre todo para las empresas que invierten en el desarrollo de una cultura de ciencia de datos.[16]

Para conseguir el éxito, los potenciales ciudadanos científicos de datos deberán contar con lo siguiente:

- acceso a los datos
- curiosidad
- facilidad con SQL
- conocimiento sobre dominios
- colaboración

RESUMEN DEL CAPÍTULO 2

1 El **almacenamiento en la nube** les permite a las organizaciones recopilar datos estructurados y no estructurados de cualquier tipo y en su formato nativo. La centralización de datos en el almacenamiento en la nube crea una base para la realización de análisis, en la que se postergan los detalles hasta que las organizaciones tengan interrogantes empresariales concretos para los que requieran sus datos.

2 Un **almacén de datos en la nube** les permite a las organizaciones recopilar datos provenientes de sistemas aislados diversos para la realización de análisis, incluidos los datos almacenados en la nube, los de las bases de datos transaccionales y de análisis in situ o en la nube, o los que están almacenados con otros servicios de nube. Las organizaciones pueden ejecutar consultas, generar informes y crear visualizaciones sin tener que administrar la infraestructura subyacente.

3 El **acceso basado en funciones** democratiza los análisis en una organización. Un almacén de datos en la nube puede comprender a toda la empresa o se puede organizar de forma flexible según la estructura de la organización.

CASO DE CLIENTE

CENTRO DE MEDICINA PERSONALIZADA DE COLORADO

El Centro de medicina personalizada de Colorado (CCPM) realiza una investigación de vanguardia a través del análisis del ADN de pacientes para predecir el riesgo de enfermedades y desarrollar tratamientos orientados según el perfil genético de una persona. CCPM usa *Health Data Compass*, su almacén de datos médicos comercial. Health Data Compass integra datos genómicos de pacientes desde CCPM e historias clínicas electrónicas de UCHealth, el Hospital de Niños de Colorado y CU Medicine, incluidos los registros externos, como reclamos de seguros, registros de salud pública y datos ambientales.

Antes, Health Data Compass usaba un sistema in situ tradicional para almacenar y analizar datos. Sin embargo, el mantenimiento de este sistema resultó ser muy costoso y no era escalable para las necesidades de análisis existentes del centro, mucho menos para su crecimiento proyectado. Después de un extenso proyecto piloto de seis meses, Health Data Compass migró a GCP y Tableau, cuya combinación permite administrar conjuntos de datos enormes y realizar potentes análisis de datos visuales, con un costo menor y con la posibilidad de escalar el sistema fácilmente a medida que crezca el CCPM. Un factor importante para la decisión de CCPM fue la capacidad de GCP (incluido BigQuery, el almacén de datos de Google Cloud) de ayudar en el cumplimiento de la HIPAA según los requisitos del CCPM.

"Nos tomamos muy en serio nuestra responsabilidad de proteger los datos de los pacientes. Google Cloud Platform proporciona ventajas significativas en la seguridad de datos sobre sistemas in situ y nos permite lograr el cumplimiento de la HIPAA", señaló Michael Ames, director asociado de Health Data Compass y director de arquitectura empresarial de CCPM.[17]

MÁS INFORMACIÓN

EMPRESA

Centro de medicina personalizada de Colorado

INDUSTRIA

Salud

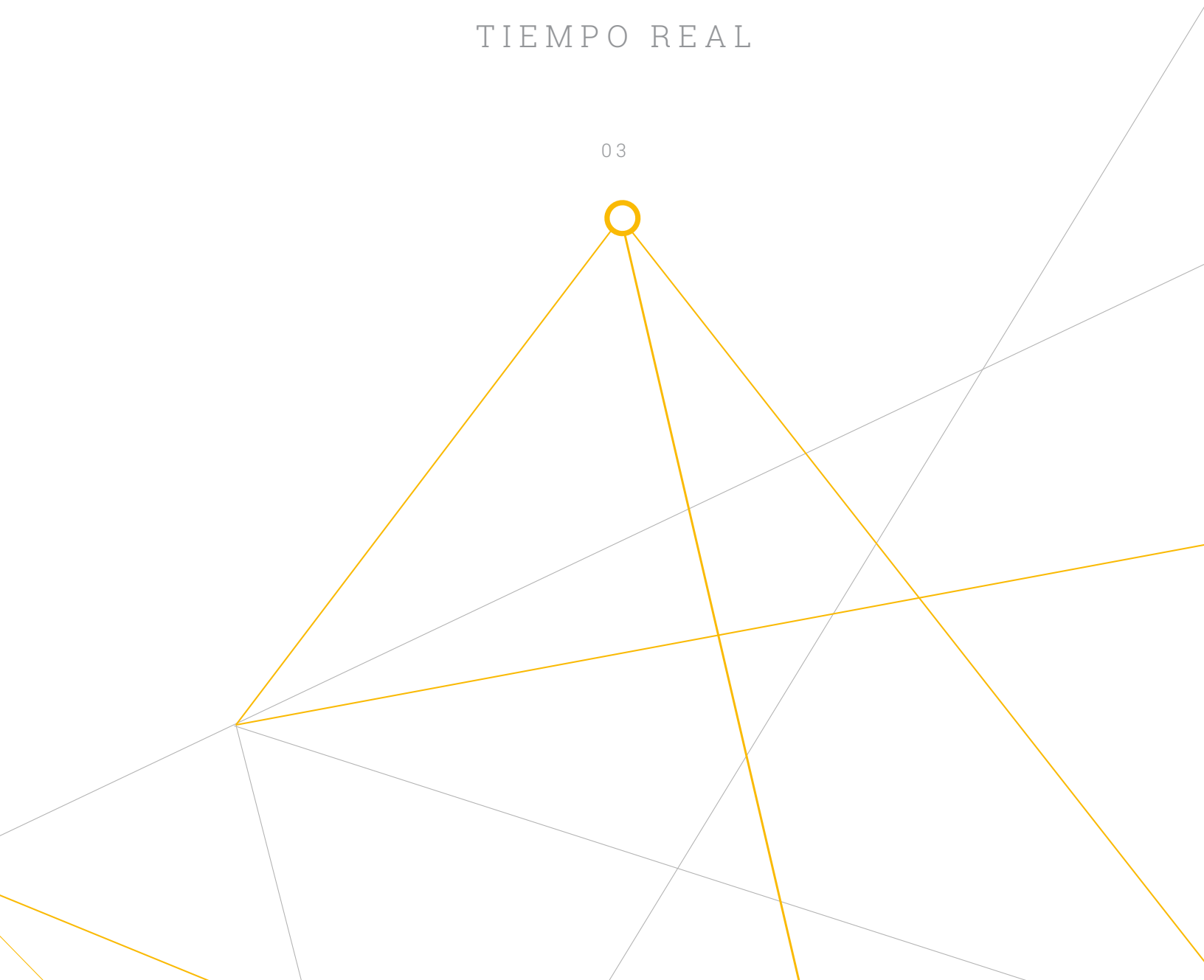
ACERCA DE

El Centro de medicina personalizada de Colorado (CCPM) es una asociación entre la Universidad de Colorado en Denver, UCHealth, el Hospital de Niños de Colorado y CU Medicine.

CAPÍTULO 3

INTEGRACIÓN DE DATOS EN TIEMPO REAL

03



INTEGRACIÓN DE DATOS EN TIEMPO REAL

03

Los científicos de datos informan que pierden entre el 50% y 80% (cifra que va en aumento) de su tiempo en dedicarse a la "organización, extracción y administración de datos" a fin de preparar los datos para el análisis.[18] La necesidad de aprovisionar recursos y escalar clústeres de servidores de forma ascendente o descendente de acuerdo con las cargas de trabajo impredecibles sigue afectando a los equipos que preparan datos in situ.[19]

Menos trabajo de "protección de datos" con servicios administrados

Los servicios de nube completamente administrados permiten aislar a TI del trabajo de infraestructura involucrado en la integración de datos y preparación de datos a gran escala. Piensa en un termostato inteligente que busca aprender y ajustarse a las preferencias de los diferentes equipos en un edifico de oficinas. Mientras el termostato está en uso, la nube se alimenta con datos de uso sin procesar, como la configuración de la temperatura y los niveles de consumo de energía a lo largo del día. A medida que se recopilan los datos, se puede acelerar una canalización de procesamiento a pedido para preparar los datos que no se han procesado: asegurar la baja de entradas dentro de un rango válido, convertir la temperatura y el uso de la energía en las unidades deseadas y formatear los datos de tiempo. La canalización de datos estructura formalmente esos datos y, luego, carga los resultados transformados en un almacén de datos en la nube. Las consultas, los informes y las visualizaciones están disponibles al instante.

de las empresas

MUESTRAN INTERÉS EN IMPLEMENTAR LA PREPARACIÓN DE DATOS DE AUTOSERVICIO PARA RESPALDAR LAS INICIATIVAS CON MACRODATOS.[20]

Con los servicios de nube completamente administrados, los recursos de infraestructura que son obligatorios para respaldar este flujo de trabajo se ubican de manera automática y, luego, se desacelera su uso. Las empresas solo pagan por los recursos que usan, lo que permite eliminar el gasto y las conjeturas en las proyecciones.

Obtención de análisis de datos en tiempo real

Aunque los sistemas tradicionales se centraban en el análisis de datos sin conexión "en lote", la demanda de estadísticas en tiempo real requiere un enfoque nuevo. Los sistemas de análisis de transmisión basados en la nube se desarrollan para administrar transmisiones de datos provenientes de aplicaciones web, smartphones o millones de sensores de IoT en tiempo real. Se pueden instalar cientos de miles de sensores en equipo en campo para informar su estado sin procesar a la nube de forma continua con el fin de realizar el procesamiento y la supervisión. Los feeds de imágenes se pueden analizar en tiempo real para las aplicaciones como la detección de anomalías y el reconocimiento facial o de objetos. Gracias a la realización de pruebas y la implementación extensa de los servicios de nube que se usan para casos de usos como estos, el análisis de datos de transmisión se puede implementar en solo unos días.

Con el *análisis de datos de transmisión en tiempo real*, los datos se transmiten directamente a la canalización de procesamiento. Luego, los datos transformados se pueden integrar en un almacén de datos en la nube, lo que permite la realización de consultas, informes y visualizaciones en segundos. De esta manera, la canalización de procesamiento sirve como un middleware que se puede acelerar a pedido, capaz de unirse a la transmisión de datos en tiempo real con datos de lote que se obtienen del almacenamiento. Los datos pueden estructurarse de manera flexible para responder los interrogantes empresariales de la organización a medida que surjan.

Por lo tanto, las organizaciones tienen dos rutas complementarias (lote y transmisión) por medio de las cuales pueden recopilar, preparar y, además, integrar datos provenientes de cualquier fuente y a cualquier destino. Los servicios de nube administrados permiten seguir ambas rutas sin complicaciones.

CÓMO APROVECHAR AL MÁXIMO TUS INVERSIONES DE MACRODATOS EXISTENTES

Muchas empresas con miras en el futuro ya usan los macrodatos, a menudo basándose en herramientas de código abierto, como Apache Hadoop y Apache Spark. Para esas empresas, es posible proteger las inversiones existentes en talento y herramientas, al mismo tiempo en que se aprovechan las ventajas de productividad de la nube.

Se extendió la adopción de herramientas de macrodatos de código abierto y sigue en aumento. Mundialmente, muchas empresas están aumentando la cantidad de datos no estructurados y están almacenándolos en sistemas de archivos en la nube pública (incluido Hadoop). Más de un tercio de los encuestados en Estados Unidos y Europa, y más de la mitad en Asia-Pacífico, informan que están implementando, implementaron o están expandiendo su implementación de Hadoop (incluidos HBASE, Accumulo, MapR Cloudera y Hortonworks). De forma similar, cerca de un tercio de los encuestados en Estados Unidos y Europa, y una gran cantidad en Asia-Pacífico (60%), están implementando, implementaron o están expandiendo su implementación de plataformas de datos en memoria (incluido Apache Spark, SAP Hana, Kognitio, Terracotta y Gigaspaces).

Para organizaciones como estas, la nube ofrece dos opciones principales:

- Seguir administrando proyectos de macrodatos con herramientas de código abierto familiares, pero migrar a máquinas virtuales en la nube. Se aplican beneficios de nube comunes: quitar bienes de capital costosos; cambiarse a un modelo de gastos operativos de facturación, en el que las organizaciones pagan según los datos almacenados y procesados; escalar sin complicaciones. Ten en cuenta que, en este modelo de montacargas, los equipos de TI y programadores aún deben administrar sus propias canalizaciones de procesamiento de datos y almacenamiento. Sin embargo, es la ruta más directa para aprovechar el talento, las herramientas y las relaciones con los proveedores ya establecidas.
- La nube ofrece versiones completamente administradas para muchas de las herramientas de código abierto más populares en macrodatos. Por ejemplo, ejecutar Apache Hadoop, Apache Spark, Apache Pig y Apache Hive en la nube disminuye las tareas de administración de datos básicas como la implementación, el registro y la supervisión.[21] Esta es una opción excelente para los equipos que buscan aprovechar los mejores aspectos tanto de los entornos nativos de nube como de los entornos in situ.

Cualquiera de estas opciones les permite a las organizaciones proteger sus inversiones existentes en la implementación de macrodatos y, al mismo tiempo, usar de manera inteligente la economía de la nube para controlar costos y obtener flexibilidad.

RESUMEN DEL CAPÍTULO 3

1 **Las canalizaciones de procesamiento de datos basadas en la nube** les permiten a las organizaciones extraer, transformar o preparar y, también, integrar datos de cualquier fuente a cualquier destino (in situ o nube).

2 **Los enfoques sin servidores** para la preparación de datos administran completamente la infraestructura subyacente. Los recursos se ubican de manera automática según las necesidades de cada canalización de procesamiento de datos.

3 **Los análisis de transmisión en la nube** permiten transmitir datos de las aplicaciones web, para dispositivos móviles y de IoT a las canalizaciones de procesamiento de datos en tiempo real. Desde allí, se pueden procesar e integrar datos en un almacén de datos en la nube para obtener una vista en tiempo real de la empresa.

CASO DE ÉXITO

CITIBANK UK

En esta prueba de concepto, la tarea del equipo era demostrar qué tan sencillo sería para Citibank usar Google BigQuery y Pub/Sub de Google Cloud a fin de analizar y consumir datos de Thomson Reuters de indicadores históricos casi en tiempo real de 1,000 instrumentos financieros. El trabajo se realizó en colaboración con Sean Micklethwaite, programador principal de Citibank, y Sebastian Fuchs, especialista de soluciones de Thomson Reuters.

"Queríamos una API que pudiésemos consultar para obtener datos históricos a pedido, sin tener que mantener nuestro propio almacén de datos ni correr con todos los costos y gastos generales operativos que conlleva", explicó Micklethwaite. "Además, exigimos actualizaciones en tiempo real de los datos de mercado con latencia a nivel humano. Con Google Cloud, obtenemos acceso a todos los datos que necesitamos en una sola plataforma. BigQuery se ocupa de nuestras necesidades de datos de indicadores históricos y cuenta con la potencia necesaria para procesar indicadores con gran frecuencia durante grandes intervalos de tiempo. Pub/Sub de Cloud se encarga de nuestras necesidades de datos en tiempo real y recibimos todos los datos en un formato uniforme".

Fuchs agregó: "Comenzamos a usar BigQuery sin tener la necesidad de realizar planificaciones sobre capacidad excesivas por adelantado. Simplemente, crece a medida que lo necesitamos, desde una [perspectiva] de aprovisionamiento de contenido, así como desde un punto de vista de cantidad de consultas de usuarios".

MÁS INFORMACIÓN

EMPRESA

Citibank UK

INDUSTRIA

Servicios financieros

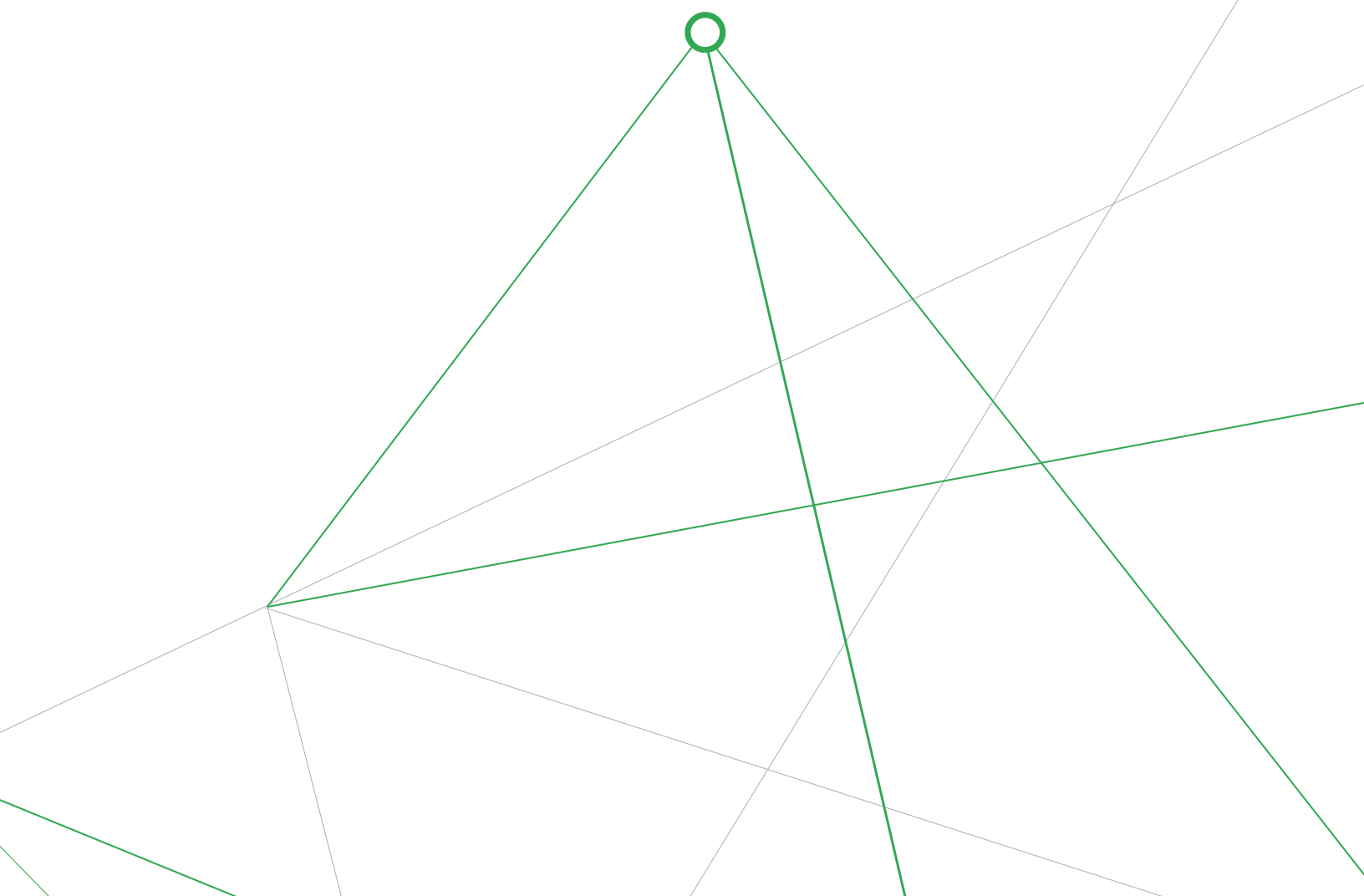
ACERCA DE

En un experimento de prueba de concepto, Google Cloud se asoció con Thomson Reuters para demostrarle al departamento de Mercados y Banca Global de Citibank los beneficios de combinar las tecnologías de datos principales de Google con el contenido de mercados financieros de Thomson Reuters.

CAPÍTULO 4

INTELIGENCIA ARTIFICIAL Y APRENDIZAJE AUTOMÁTICO

04



INTELIGENCIA ARTIFICIAL Y APRENDIZAJE AUTOMÁTICO

04

Con frecuencia, los avances recientes en el aprendizaje automático (ML) y la inteligencia artificial (IA) aparecen en los titulares de los periódicos. Las computadoras superaron a los humanos campeones en Go, un juego de mesa con más posiciones que átomos en el universo.[22] Perfeccionaron los videojuegos populares y, fundamentalmente, aprendieron a reconocer gatos.[23] Hace poco, un trabajo en IA logró grandes ahorros en costos de energía y se destacó al aprendizaje automático como "marco de trabajo para fines generales que permite comprender dinámicas complejas".[24] Este marco de trabajo está comenzando a aplicarse de varias maneras y a proporcionar resultados en muchas industrias.

El concepto de la IA es simple: es la capacidad del software de mejorar sin tener que programarse de manera explícita. En lugar de requerir que los programadores escriban código nuevo de forma manual, la IA se basa en algoritmos capaces de volverse "más inteligentes" mediante el análisis de más datos del mundo real. Centralizar el almacenamiento y la preparación de datos en la nube (los objetivos del segundo y tercer capítulo respectivamente) crea la base ideal para la capacitación y mejora de los modelos de IA.

Las oportunidades que ofrece la IA van más allá de solo automatizar las tareas que solían ser manuales. En la venta minorista en línea, por ejemplo, los algoritmos de aprendizaje automático pueden transferir y analizar enormes cantidades de datos de consumidores a medida que los compradores potenciales navegan en la tienda en línea o la app para dispositivos móviles del vendedor minorista. Mientras más datos analiza el modelo, más cerca está de comprender cuándo y por qué un comprador en particular decidirá realizar una compra determinada. Finalmente, este aprendizaje se vuelve predictivo, lo que le permite al vendedor minorista ofrecer el producto adecuado para una persona determinada en el tiempo preciso. Este nivel de personalización, que antes representaba el comerciante de un pueblo pequeño que conocía los nombres y cumpleaños de los hijos de sus clientes, ahora es posible llevarlo a cabo a escala.

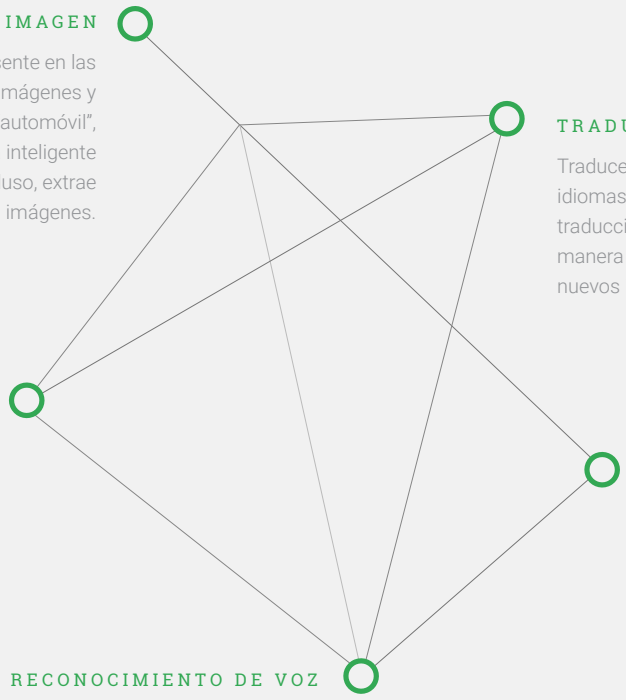
de los encuestados

CREE QUE EL ÉXITO FUTURO DE SU ORGANIZACIÓN DEPENDE DE LA IMPLEMENTACIÓN EXITOSA DEL APRENDIZAJE AUTOMÁTICO.[25]

Muchas empresas pequeñas con tecnología avanzada ya están viendo resultados con el aprendizaje automático, pero las empresas más establecidas tienen una oportunidad única de aprovechar la abundancia de datos históricos.[26] Con el aprendizaje automático, los resultados dependen de la gran cantidad de datos disponibles para alimentar los modelos de capacitación (**consulta "Modelos capacitados previamente: Un primer paso para trabajar con IA"**). Las empresas establecidas pueden aprovechar sus datos de origen, desde registros en sistemas de TI y transacciones financieras hasta transcripciones de llamados al servicio de Atención al cliente, a fin de capacitar y optimizar estos modelos, y obtener estadísticas únicas para la empresa.

MODELOS CAPACITADOS PREVIAMENTE: UN PRIMER PASO PARA TRABAJAR CON IA

La manera más directa de comenzar a trabajar con la IA es usar modelos de aprendizaje automático capacitados previamente, disponibles de inmediato a través de la nube. No se requieren conocimientos previos sobre el aprendizaje automático. Es posible que estas capacidades les resulten conocidas a las personas que usan aplicaciones populares para consumidores, en las que algunos modelos alcanzan niveles de precisión predictiva que excede la capacidad humana:



ANÁLISIS DE IMAGEN

Comprende el contenido presente en las imágenes. Para ello, proporciona imágenes y obtén categorías de texto (p. ej., "automóvil", "Torre Eiffel"). Detecta de manera inteligente rostros y objetos individuales o, incluso, extrae palabras impresas presentes en las imágenes.

TRADUCCIÓN

Traduce todo el texto a uno de los tantos idiomas disponibles. Los servicios de traducción mejoran y se expanden de manera continua para comprender nuevos idiomas.

ANÁLISIS DE VIDEO

Identifica objetos y entidades significativas presentes en los videos. Comprende lo que ocurre a nivel de una escena o un fotograma.

ANÁLISIS DE TEXTO

Comprende la estructura y el significado semántico del texto. Extrae información sobre las personas, los lugares y los eventos presentes en cualquier texto. Comprende sentimientos y analiza la información presente en los artículos, los blogs, las conversaciones o registros de chats con Atención al cliente, o cualquier texto nuevo que proporciones.

RECONOCIMIENTO DE VOZ

Convierte audio en texto con una precisión extraordinaria. Proporciona audio a la API y obtén el texto completo al instante, con compatibilidad para muchos idiomas. Permite casos de uso en tiempo real, como la transcripción, o crea aplicaciones que se dirigen por voz.

Estos servicios son generales (no se limitan a las aplicaciones para consumidores) y se pueden integrar fácilmente en cualquier aplicación mediante simples llamadas de API. No es necesario que los programadores conozcan ninguno de los detalles subyacentes. Sin tener que desarrollar ninguno de estos servicios de manera interna, las empresas pueden acceder a las capacidades más recientes al instante, como un servicio.

Por lo general, las empresas establecidas y sus semejantes de la industria tienen décadas de datos de origen acumulados: de transacciones financieras, registros de sistemas, datos sin procesar generados por fabricantes, ventas minoristas y datos de comercio electrónico recopilado durante años, y resultados del rendimiento de las campañas de marketing. Estos datos, refinados y usados adecuadamente para capacitar modelos de aprendizaje automático personalizados, se convierten en una fuente de poder predictivo. Las empresas establecidas, en lugar de reutilizar los servicios predeterminados, pueden usar los datos de origen a fin de optimizar los procesos empresariales para sus clientes, que son una fuente poderosa de diferenciación.

Los casos de uso incluyen a muchas industrias y revelan algunas de las aplicaciones de IA más prometedoras. La detección de fraudes en los servicios financieros y el mantenimiento preventivo en el sector de la fabricación resaltan la capacidad de identificar anomalías en una enorme cantidad de transacciones y registros desordenados, una necesidad común en muchos dominios. Las sugerencias de tratamiento y diagnóstico en el sector de la salud y de juicios sobre solvencia resaltan la capacidad del aprendizaje automático de brindar asistencia con categorización, que por lo general también es muy útil.

CUANTIFICACIÓN DE LA GANANCIA

En asociación con la empresa de investigaciones M-Brain, Google Cloud encuestó a 20 líderes empresariales y de TI que implementaron proyectos de aprendizaje automático sobre los beneficios clave obtenidos de los proyectos. Estos son algunos de los principales beneficios que se mencionaron:

- ahorro de tiempo
- ahorros de costos
- gestión de riesgos mejorada
- calidad de análisis mejorada
- mayores ingresos

Otros mencionaron la automatización, el servicio mejorado y la planificación de inventario mejorada. [27]

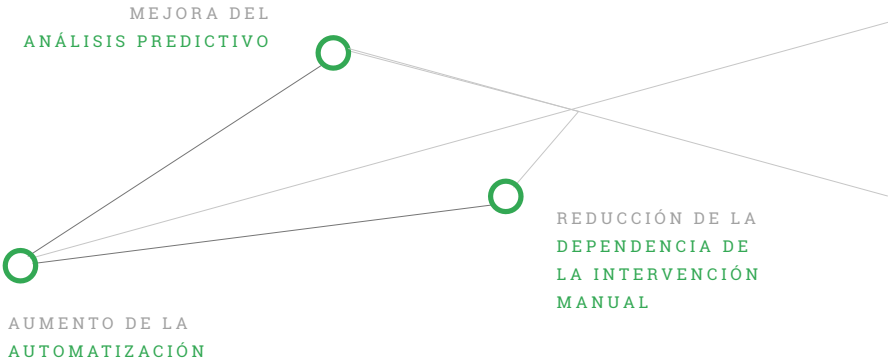
Círculo virtuoso: Recopilar, preparar, capacitar y predecir

Las capacidades que se presentaron en el segundo y tercer capítulo funcionan como una base para los modelos de aprendizaje automático de capacitación que usan datos de origen. Con los datos sin procesar ya centralizados en el almacenamiento en la nube y en un almacén de datos en la nube, las canalizaciones de datos sin servidores pueden extraer continuamente estos datos y prepararlos para capacitar modelos de aprendizaje automático personalizados. Debido a que los modelos de aprendizaje automático pueden almacenarse en la nube por sí solos, se encuentran disponibles de inmediato para que las aplicaciones hagan predicciones. Este bucle forma un círculo virtuoso, en el que los modelos de aprendizaje automático almacenados en la nube siguen mejorando gracias a los datos nuevos de capacitación, lo que mantiene los modelos actualizados y relevantes.

LAS NECESIDADES EMPRESARIALES PRINCIPALES QUE SE INFORMARON

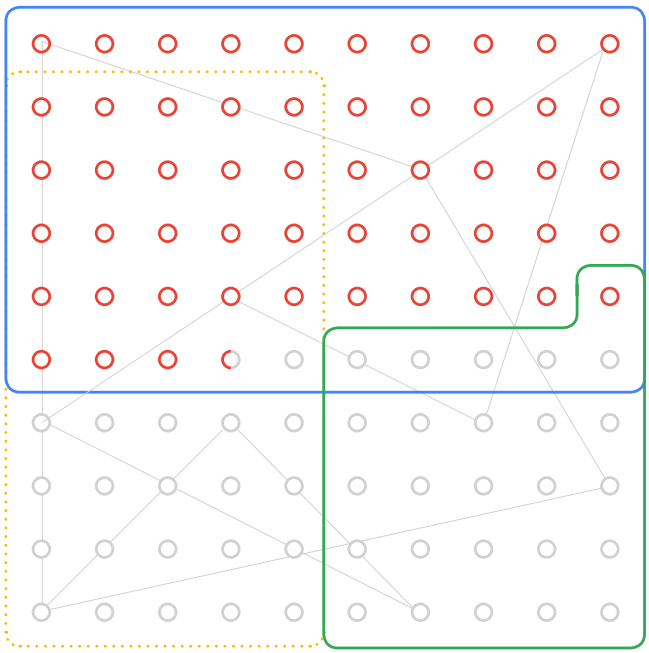
en las siguientes áreas:[28]

- SALUD
- SERVICIOS FINANCIEROS
- SECTOR DE FABRICACIÓN
- VENTA MINORISTA
- MEDIOS DE COMUNICACIÓN/ VIDEOJUEGOS



APRENDIZAJE AUTOMÁTICO: EL NUEVO CAMPO DE PRUEBAS PARA LA VENTAJA DE LA COMPETITIVIDAD

La era del aprendizaje automático ha llegado finalmente y ya se encuentra en pleno apogeo en las empresas más pequeñas que cuentan con tecnología avanzada, según una nueva encuesta de líderes tecnológicos y empresariales realizada por *MIT Technology Review* Custom. Estos son algunos de los resultados clave:[29]



60 % de los encuestados ya **implementaron estrategias de aprendizaje automático.**

MÁS DEL 50 % de **los encuestados que implementaron el aprendizaje automático y se encuentran en la primera fase, ya están observando el ROI.**

45 % ha **logrado obtener estadísticas y análisis de datos más extensos.**

26% informa una ventaja **de competencia mayor.**

Descarga el informe completo aquí.

RESUMEN DEL CAPÍTULO 4

1. La IA y su subconjunto de aprendizaje automático son simples por concepto: la capacidad que tiene el software de mejorar **sin tener que programarlo de manera explícita**.
2. La IA depende de grandes cantidades de datos de prueba, lo que les ofrece a las empresas establecidas la **ventaja única** de extraer grandes cantidades de datos empresariales generados durante largos períodos de operación.
3. El almacenamiento en la nube, el almacén de datos, la integración de datos y los análisis **proporcionan una base natural** de IA y aprendizaje automático. Para ello, ponen los datos a disposición para la capacitación y la optimización en tiempo real, lo que potencia un círculo virtuoso de mejora continua.

CONCLUSIÓN

En una época de abundancia de datos y respuestas inmediatas, la capacidad de extraer valor de ellos, independientemente de la fuente, el tamaño y los requisitos relacionados con la rapidez, será el centro de la ventaja competitiva de una organización.

El primer paso es volver a reformular una estrategia de datos. Las herramientas de la nube actuales les permiten a las empresas administrar enormes cantidades de tipos de datos diversos de manera más eficaz y a más bajo costo que antes. Las empresas que cuentan con un enfoque moderno para recopilar, almacenar, preparar y analizar sus datos contarán con la base para aprovechar el aprendizaje automático y la inteligencia artificial. Finalmente, estas capacidades nuevas generarán relaciones más cercanas entre las empresas y sus clientes, lo que les permitirá ser más predictivas en cada interacción.

OBTÉN MÁS INFORMACIÓN SOBRE LO QUE GOOGLE CLOUD PUEDE HACER POR TU EMPRESA.

Almacenamiento y bases de datos

MÁS INFORMACIÓN

Soluciones para macrodatos

MÁS INFORMACIÓN

Inteligencia artificial y aprendizaje automático

MÁS INFORMACIÓN

TRABAJOS CITADOS

1. Un 81% de ejecutivos séniores encuestados por Ernst & Young acordaron que los datos deberían ser el centro de la toma de decisiones, solo el 31% había reestructurado significativamente las operaciones para incorporar macrodatos y un 23% había implementado estrategias de datos a nivel de empresa. Ernst & Young, *Becoming an Analytics-Driven Organization* (2015) (vínculo).
2. David Reinsel et al., *Data Age 2025: The Evolution of Data to Life-Critical* (IDC, 2017) (vínculo).
3. Cade Metz, "Exclusive: Inside Google Spanner, the Largest Single Database on Earth", *Wired* (26 de noviembre de 2012) (vínculo).

   Cade Metz, "Spanner, the Google Database that Measured Time, Is Now Open to Everyone", *Wired* (14 de febrero de 2017) (vínculo).
4. Robert McMillan, "Inside the Artificial Brain that's Remaking the Google Empire", *Wired* (16 de julio de 2014) (vínculo).TensorFlow (vínculo).
5. Forrester, *Forrester's Global Business Technographics Data and Analytics Survey* (2016) (vínculo).
6. Leandro DalleMule and Thomas H. Davenport, "What's Your Data Strategy?" *Harvard Business Review* (mayo de 2017) (vínculo).
7. John Gantz and David Reinsel, *The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East* (IDC, 2012) (vínculo).
8. Tracie Kambies et al., *Tech Trends 2017: Dark Analytics: Illuminating Opportunities Hidden within Unstructured Data* (Deloitte University Press, 2017) (vínculo).
9. *Precios de Google Cloud Storage*, Google Cloud Platform (vínculo).
10. Forrester, *Forrester's Global Business Technographics Data and Analytics Survey* (2016) (vínculo).
11. "Three Ways Marketing Organizations Can Make Data More Actionable", *Harvard Business Review* (9 de agosto de 2016) (vínculo).
12. Los almacenes de datos modernos en la nube respaldan la importación (incluso la realización de consultas adhoc) de muchos formatos semiestructurados de manera automática. Para datos no estructurados que primero deben transformarse (p. ej., ETL), consulta el **capítulo 3: Preparación de datos**.
13. Forrester, *Forrester's Global Business Technographics Data and Analytics Survey* (2016) (vínculo).
14. Forrester, *Forrester's Global Business Technographics Data and Analytics Survey* (2016) (vínculo).
15. James Manyika et al., *Big Data: The Next Frontier for Innovation, Competition, and Productivity* (Instituto Global McKinsey, 2011) (vínculo).
16. Lisa Morgan, "Citizen Data Scientists: 7 Ways to Harness Talent", *InformationWeek* (24 de julio de 2015) (vínculo).
17. *Colorado Center for Personalized Medicine: Mejorar la salud mediante la integración de los registros de pacientes con información genética en Google Cloud Platform y Tableau* (Google Cloud Platform, 2017) (vínculo).
18. Steve Lohr, "For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights", *New York Times* (17 de agosto de 2014) (vínculo).
19. Forrester, *Forrester's Global Business Technographics Data and Analytics Survey* (2016) (vínculo).
20. Forrester, *Forrester's Global Business Technographics Data and Analytics Survey* (2016) (vínculo).
21. *Apache Hadoop*, The Apache Software Foundation (vínculo).

    Apache Spark, The Apache Software Foundation (vínculo).

    Apache Pig, The Apache Software Foundation (vínculo).

    Apache Hive, The Apache Software Foundation (vínculo).
22. Paul Mozur, "Google's A.I. Program Rattles Chinese Go-Master As It Wins Match", *New York Times* (25 de mayo de 2017) (vínculo).
23. Nicola Twilley, "Artificial Intelligence Goes to the Arcade", *The New Yorker* (25 de febrero de 2015) (vínculo).

    John Markoff, "How Many Computers to Identify A Cat? 16,000", *The New Yorker* (25 de junio de 2012) (vínculo).
24. James Vincent, "Google Uses DeepMind AI to Cut Data Center Energy Bills", *The Verge* (21 de julio de 2016) (vínculo).
25. *Harvard Business Review Analytic Services Global Data and Analytics Survey*, patrocinada por Google (2017).

26. Una encuesta realizada por *MIT Technology Review* demuestra que las organizaciones más pequeñas están bien encaminadas en la adopción del aprendizaje automático y sus beneficios: 60% de un total de 375 encuestados, de los que casi dos tercios eran empresas con menos de 1,000 empleados basadas en gran parte en las industrias de servicios financieros, tecnología y empresas. *MIT Technology Review* Custom and Google Cloud, *Machine Learning: The New Proving Ground for Competitive Advantage* (2017) (vínculo).
27. Anna Rader, *Machine Learning Initiatives Across Industries: Practical Lessons from IT Executives* (M-Brain, patrocinado por Google, 2017) (vínculo).
28. Anna Rader and Irida Jano, *Machine Learning Market Research: How Leading Industries Are Adopting AI* (M-Brain 2017) (vínculo).
29. *MIT Technology Review* Custom and Google Cloud, *Machine Learning: The New Proving Ground for Competitive Advantage* (2017) (vínculo).