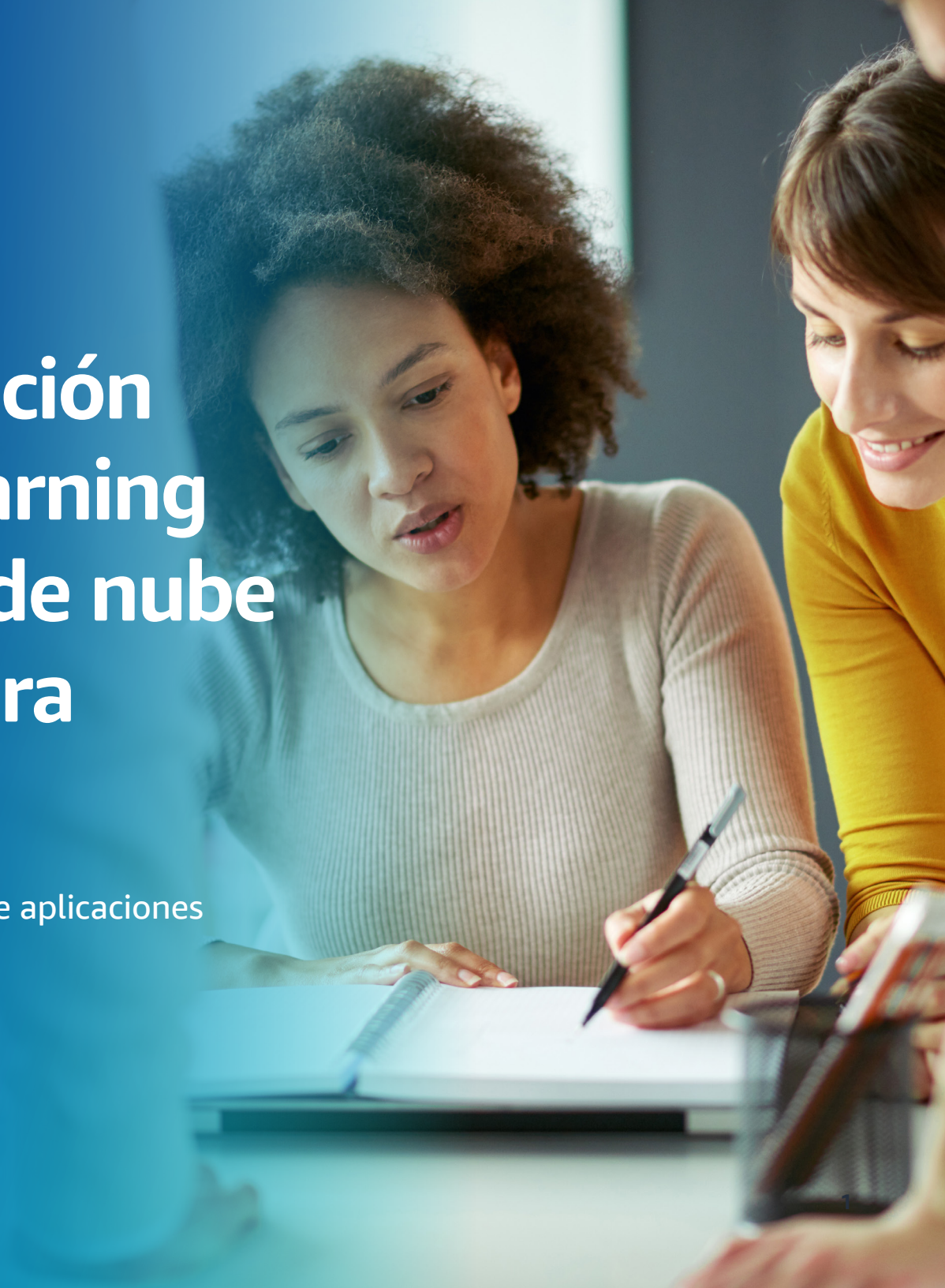
Acelere la innovación en el machine learning con los servicios de nube y la infraestructura adecuados

Prepare datos y cree, forme e implemente aplicaciones de machine learning con facilidad



Índice

Innove con el machine learning

Gracias a los avances de la capacidad informática, la disminución del precio del almacenamiento y la prevalencia de la informática en la nube, la inteligencia artificial (IA) y el machine learning (ML) han pasado a ser la tendencia predominante. Organizaciones e industrias de diferentes tamaños —incluidas las de finanzas, comercio minorista, moda, bienes inmuebles, atención de la salud y muchas más— pueden aprovechar la IA y el ML para ofrecer una amplia gama de beneficios empresariales. Entre ellos, la adquisición de nuevos y más amplios conocimientos sobre los clientes, la identificación de amenazas cibernéticas y la respuesta a estas, la toma de decisiones más inteligentes basadas en datos y la mejora de los procesos de contratación.

Gracias a estos beneficios, cada vez más organizaciones están invirtiendo en IA y ML. IDC prevé que el gasto mundial en IA y ML aumentará a una tasa compuesta de crecimiento anual (CAGR) del 26,5 % de 2022 a 2026, pasando de 118 000 millones de USD en 2022 a más de 300 000 millones de USD en 2026.[1]

Los modelos de ML se utilizan para muchos casos de uso, como el procesamiento de lenguaje natural (NLP), la visión artificial (CV) y el procesamiento de documentos, que aprenden de los datos existentes mediante un proceso denominado entrenamiento para tomar decisiones sobre nuevos datos a través de un proceso denominado inferencia. Los modelos de ML a gran escala actuales, conocidos como modelos fundacionales (FM), contienen cientos de miles de millones de parámetros, que son modelos de propósito general increíblemente poderosos que pueden ampliarse y personalizarse para casos de uso específicos, sin tener que construir un modelo desde cero cada vez. Estos modelos potencian las aplicaciones de IA generativa, como el resumen de textos, la generación de códigos y la creación de imágenes, a partir de indicaciones en lenguaje natural.

Algunos de los algoritmos más populares actualmente son los siguientes:

- **Procesamiento de lenguaje natural (NLP)**: los algoritmos NLP analizan el lenguaje a gran escala y tienen la capacidad de entender el contexto, analizar el habla y realizar traducciones casi en tiempo real. Se utilizan para crear aplicaciones de ML, como chatbots, filtros de spam, asistentes de voz y herramientas de monitoreo de redes sociales.
- **Visión artificial (CV)**: los algoritmos de visión artificial procesan y analizan datos visuales para detectar objetos y clasificar imágenes de manera similar a como lo hace la mente humana, pero a una velocidad y escala exponencialmente mayores. Se pueden utilizar para mejorar la seguridad en el lugar de trabajo, admitir la verificación de identidad de manera digital y señalar los contenidos inapropiados.
- **Procesamiento de documentos**: los algoritmos de procesamiento de documentos extraen texto, escritura a mano y datos de los documentos, y van más allá del reconocimiento óptico de caracteres (OCR) para identificar y entender datos de formularios y tablas. Pueden utilizarse para extraer información de historias clínicas y automatizar el procesamiento de documentos financieros.
- **IA generativa**: los FM, como los modelos de lenguaje de gran tamaño (LLM) y los modelos de difusión, pueden utilizarse para generar contenidos originales similares a los humanos, como prosa coherente, imágenes y videos basados en indicaciones de lenguaje natural. Estos pueden utilizarse para aplicaciones como la generación de códigos, el resumen de textos, las respuestas a preguntas y la generación de imágenes y videos.

[1] "El gasto mundial en sistemas centrados en la IA superará los 300 000 millones de USD en 2026, según IDC", IDC, septiembre de 2022

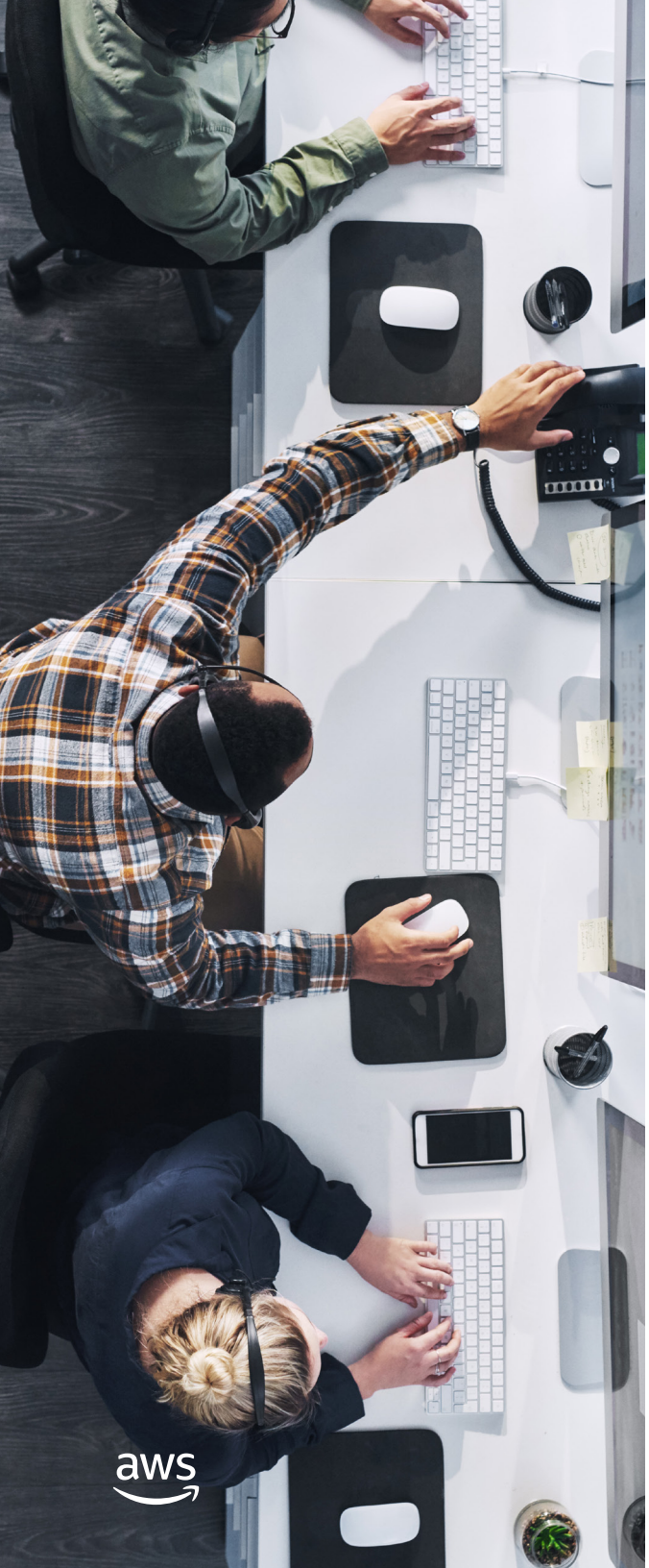
El valor empresarial potencial de estas aplicaciones es considerable, pero también lo son los requisitos de recursos e infraestructura necesarios para hacerlas funcionar a velocidad y a gran escala. El entrenamiento de modelos de ML que impulsa estos casos de uso requiere grandes cantidades de datos, decenas de miles de nodos de computación y redes mejoradas inter e intranodos.

En respuesta a estos requisitos, cada vez más organizaciones recurren a la nube. La nube combina los datos, el almacenamiento a bajo costo, la seguridad y los servicios de ML con infraestructura de computación de alto rendimiento (HPC) para el entrenamiento y la implementación de modelos.

Cómo AWS acelera el éxito del machine learning

En Amazon Web Services (AWS), se desarrolla más ML que en ningún otro sitio, y AWS ofrece la cartera de servicios más amplia y profunda para acelerar la transformación empresarial. Organizaciones de todos los tamaños, desde las empresas de Fortune 500 hasta las startups, se están beneficiando de la combinación ideal de infraestructura y servicios de ML de alto rendimiento y bajo costo de AWS. Al ejecutar sus cargas de trabajo de ML en la nube, los clientes obtienen acceso bajo demanda a la infraestructura y las herramientas de ML que pueden ponerse en marcha en minutos, pueden escalar de una a miles de instancias y solo pagan por lo que utilizan.

Veamos algunos ejemplos de clientes de AWS que actualmente obtienen resultados con ML.



Logre el éxito con AWS Machine Learning

Decenas de miles de clientes han elegido AWS ML como herramienta para obtener una gran variedad de resultados empresariales. Estos son algunos ejemplos:

- **LG AI Research** desarrolló EXAONE, un FM que contiene 300 000 millones de parámetros. EXAONE se creó mediante **Amazon SageMaker** para realizar una amplia gama de tareas en diferentes sectores, como la moda, la fabricación, la investigación, la educación y las finanzas. Con ayuda del FM, desarrollaron una IA llamada Tilda, que colaboró con un diseñador de moda para generar 3000 imágenes y patrones con el fin de diseñar más de 200 conjuntos para la Semana de la Moda de Nueva York de 2022. Gracias a SageMaker, LG AI Research redujo sus costos en alrededor del 35 % y aumentó la velocidad de procesamiento de datos en un 60 %.
- **NerdWallet** proporciona herramientas y consejos que facilitan a los clientes la administración de sus finanzas. La empresa confía plenamente en la ciencia de datos y el ML para conectar a los clientes con productos financieros personalizados. NerdWallet utiliza una serie de servicios de AWS, como SageMaker y las **instancias P3 de Amazon Elastic Compute Cloud (Amazon EC2) P3**, para mejorar el rendimiento y reducir el tiempo que requieren los científicos de datos para entrenar e iterar los modelos de ML de meses a tan solo días.
- **Sprinklr** ofrece una plataforma unificada de administración de la experiencia del cliente (Unified-CXM) que combina diferentes aplicaciones de marketing, publicidad, investigación, atención al cliente, ventas y participación en redes sociales. Esta plataforma de Sprinklr utiliza algoritmos de ML en datos no estructurados procedentes de muchos canales diferentes para ofrecer a sus clientes información sobre sentimientos e intenciones. Por ejemplo, los modelos NLP y CV ML de la empresa analizan diferentes formatos de datos procedentes de publicaciones en redes sociales, blogs, videos y otros contenidos disponibles en dominios públicos a través de más de 30 canales. Con las **instancias Inf1 de Amazon EC2**, que utilizan la tecnología de **AWS Inferentia**, un acelerador de inferencia de ML de alto rendimiento, Sprinklr pudo reducir la latencia en un 30 %. Empezar fue fácil, y ahora el equipo es capaz de implementar un modelo utilizando instancias Inf1 de Amazon EC2 en menos de dos semanas.

Acelere cada paso del ciclo de vida del machine learning

Las empresas recurren a AWS para derribar las barreras en cada paso del ciclo de vida del ML. Hay cuatro pasos importantes en el ciclo de vida del ML. En cada paso, los desarrolladores de ML necesitan apoyar la gobernanza de ML, por ejemplo, creando políticas y estableciendo controles para garantizar la transparencia del modelo, la privacidad de los datos y la seguridad.

1. Los equipos de ciencia de datos tienen que preparar datos de ejemplo para formar un modelo.
2. Después, deben seleccionar el algoritmo o el marco que utilizarán para crear el modelo.
3. A continuación, los modelos deben entrenarse para que realicen predicciones y ajustarse con frecuencia para lograr la máxima precisión.
4. Por último, los modelos se deben implementar, integrar con sus aplicaciones, monitorear, escalar y administrar en producción.

AWS le ofrece la posibilidad de elegir la infraestructura en cada paso del flujo de trabajo del ML. Puede personalizar su infraestructura, incluyendo la computación, las redes y el almacenamiento, para adaptarla a sus necesidades de rendimiento y presupuesto. Tiene una amplia gama de opciones para tener una infraestructura de alto rendimiento, rentable y escalable.

AWS ofrece la infraestructura de ML de mayor rendimiento impulsada por GPU y aceleradores de ML **AWS Trainium** y **AWS Inferentia** personalizados. AWS Trainium permite ahorrar hasta un 50 % en costos de entrenamiento con respecto a instancias comparables de Amazon EC2. AWS Inferentia2 posibilita un rendimiento de precios hasta un 70 % superior al de instancias comparables de Amazon EC2.

La forma más fácil y rápida de utilizar la **infraestructura de ML de AWS** es a través de SageMaker, un servicio totalmente administrado que agrupa un conjunto amplio de capacidades, como el etiquetado de datos, la preparación de datos, la ingeniería de características, la detección de sesgos estadísticos, el machine learning automático (AutoML), el entrenamiento, el ajuste, el alojamiento, la explicabilidad, el monitoreo y los flujos de trabajo. **Amazon SageMaker JumpStart** ofrece cientos de algoritmos integrados, FM preentrenados y soluciones prediseñadas que los clientes pueden implementar con tan solo unos clics.

El **SDK AWS Neuron** también facilita la extracción de todo el rendimiento de los aceleradores AWS Trainium y AWS Inferentia mediante la integración nativa con marcos de ML populares, como PyTorch y TensorFlow. Los clientes pueden seguir utilizando sus marcos de trabajo y código de aplicación existentes al utilizar las instancias Trn1n, Trn1, Inf2 e Inf1 de Amazon EC2 basadas en estos aceleradores.

Los clientes también pueden utilizar **AWS Deep Learning Containers** (imágenes de Docker preinstaladas con marcos de deep learning) con **Amazon Elastic Kubernetes Service** (Amazon EKS) y **Amazon Elastic Container Service** (Amazon ECS). Además, las **AMI de aprendizaje profundo de AWS** (DLAMI) ofrecen entornos preconfigurados para crear aplicaciones de deep learning al proporcionarles a los profesionales e investigadores del ML la infraestructura y las herramientas necesarias para acelerar el deep learning en la nube a cualquier escala.

Ahora que tiene una idea general de cómo funciona el proceso de desarrollo de ML y de cómo AWS puede ayudar, profundicemos en cada una de las cuatro etapas con más detalle.

PASO 1

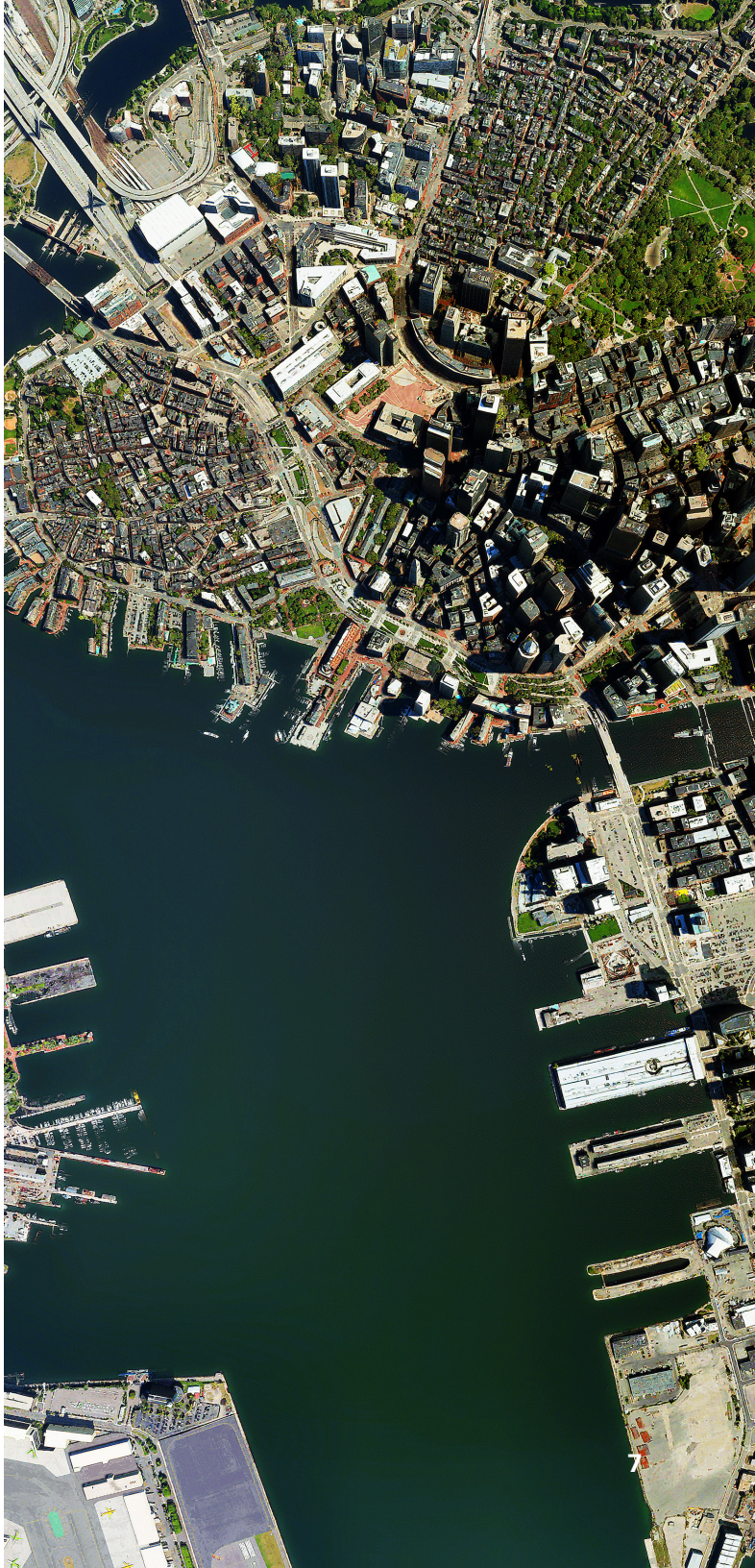
Prepare datos rápida y fácilmente

Desafíos

Los datos son el combustible del ML. Sin embargo, incluso con la estrategia de datos correcta y definida, la administración de datos puede ser la parte que demande más tiempo y presente más desafíos a la hora de crear modelos de ML. Muchos clientes dicen que pasan alrededor del 80 % de su tiempo en tareas de preparación de datos, como la recopilación, la limpieza y el etiquetado de datos.

Existen dos tipos de datos: los estructurados y los no estructurados. Los datos estructurados son datos cuantitativos sumamente organizados que son fáciles de descifrar por el ML. No obstante, los datos estructurados conforman solo un pequeño porcentaje de la totalidad de los datos. Los datos no estructurados son cualitativos e incluyen cosas como imágenes, notas escritas a mano y datos geoespaciales. Son extremadamente valiosos, pero mucho más difíciles de usar para el ML. La mayor parte de la información necesaria para el ML está contenida en datos no estructurados, pero el análisis de datos no estructurados suele estar fuera del alcance de muchas de las herramientas de administración de datos existentes; por ejemplo, cuando un médico necesita analizar información de radiografías, resonancias magnéticas y recetas escritas.

Para complicar aún más las cosas, la mayoría de los equipos de ingeniería de ML deben escribir código para las tareas comunes de preparación de datos necesarias para el ML, o integrarlas con marcos independientes de extracción, transformación y carga (ETL) que son administrados por otras organizaciones.



Solución

SageMaker ayuda a procesar tanto los datos estructurados como los no estructurados. **Amazon SageMaker Ground Truth Plus** ayuda a los clientes a crear con facilidad conjuntos de datos de entrenamiento de alta calidad sin tener que crear aplicaciones de etiquetado o administrar fuerzas de trabajo de etiquetado. SageMaker Ground Truth Plus también ayuda a reducir los costos de etiquetado de datos hasta en un 40 % y a cumplir sus requisitos de seguridad, privacidad y conformidad de datos. Solo tiene que cargar sus datos, SageMaker Ground Truth Plus se encargará de crear flujos de trabajo de etiquetado de datos y administrar los flujos de trabajo. En el caso de los datos geoespaciales, los profesionales del ML pueden acceder a orígenes de datos geoespaciales, operaciones de procesamiento personalizadas, modelos de ML preentrenados y herramientas de visualización integradas para ejecutar el ML geoespacial más rápido y a escala.

Para datos estructurados, **Amazon SageMaker Data Wrangler** simplifica drásticamente la preparación de datos estructurados con una interfaz visual sin códigos. SageMaker Data Wrangler contiene más de 300 transformaciones de datos integradas para que pueda normalizar, transformar y combinar características rápidamente sin tener que escribir ningún código. Con las plantillas de visualización de SageMaker Data Wrangler, puede previsualizar e inspeccionar rápidamente si estas transformaciones se han completado según lo previsto visualizándolas en **Amazon SageMaker Studio**, el primer entorno de desarrollo totalmente integrado (IDE) para ML. También puede simplificar sus flujos de trabajo de datos con un entorno de bloc de notas unificado para ingeniería de datos, análisis y ML. Cree, explore y conéctese a clústeres de **Amazon EMR** y sesiones interactivas de AWS Glue directamente desde los blocs de notas de SageMaker Studio. Monitoree y depure los trabajos de Spark utilizando herramientas conocidas, como Spark UI, directamente desde los blocs de notas. Utilice la capacidad integrada de preparación de datos de SageMaker Data Wrangler directamente desde los blocs de notas para visualizar los datos, identificar los problemas de calidad de los datos y aplicar las soluciones recomendadas para mejorar la calidad de los datos y la precisión del modelo sin escribir una sola línea de código.

Una vez preparados los datos, puede crear flujos de trabajo de ML totalmente automatizados con **Canalizaciones de Amazon SageMaker** y guardarlos para reutilizarlos en el **Almacén de características de Amazon SageMaker**.

> **"Con Amazon SageMaker Data Wrangler, ahora podemos seleccionar, limpiar, explorar y comprender nuestros datos de manera interactiva y eficaz, lo que permite a nuestro equipo de ciencia de datos crear canalizaciones de ingeniería de características que se pueden escalar sin esfuerzo a conjuntos de datos que abarcan cientos de millones de filas... con Amazon Sagemaker Data Wrangler, podemos poner en funcionamiento nuestros flujos de trabajo de ML de forma más rápida".**[2]
>
> **Caleb Wilkinson**, Lead Data Scientist, INVISTA

[2] Amazon SageMaker Data Wrangler, marzo de 2023

PASO 2

Cree modelos precisos a través de varios marcos

Desafíos

Una vez que cuente con los datos para el entrenamiento, deberá elegir un algoritmo de ML con un estilo de aprendizaje que se ajuste a sus necesidades. Esto puede ser difícil, ya que hay docenas de algoritmos para elegir. Los marcos de ML, como PyTorch y TensorFlow, facilitan el desarrollo, pero suelen ser más adecuados para algoritmos específicos. Con frecuencia, esto da lugar a la necesidad de administrar y crear a través de una combinación de algoritmos y marcos, lo que puede ser complejo, propenso a los errores y requerir una gran cantidad de recursos.

Para crear modelos, también se necesita mucha experimentación e iteración. La mayoría de los equipos utilizan Jupyter Notebooks para crear modelos y compartir el trabajo entre equipos. Lamentablemente, a medida que se desarrollan más modelos, se hace más difícil escalar y compartir el trabajo.

Solución

Si desea utilizar algoritmos creados previamente y un servicio totalmente administrado para crear modelos de ML eficientes, precisos y potentes, SageMaker es su solución. SageMaker incluye una docena de algoritmos creados previamente que pueden implementarse en el marco de su elección. Con SageMaker Studio, puede crear modelos en una interfaz visual única, lo que puede mejorar hasta diez veces la productividad del equipo de ciencia de datos.[3]

Amazon SageMaker Studio le da acceso, control y visibilidad total mientras entrena a su modelo. Puede cargar rápidamente los datos, crear blocs de notas nuevos y ajustar los experimentos de ML. Dentro de SageMaker Studio, se pueden realizar todas las actividades de desarrollo de ML, como blocs de notas, administración de experimentos, creación automática de modelos, depuración y detección de desviaciones de modelos y datos.

Los blocs de notas de Amazon SageMaker Studio administran instancias de computación para ver, ejecutar o compartir un bloc de notas. Los recursos de computación subyacentes son totalmente elásticos, por lo que puede aumentar o disminuir fácilmente los recursos disponibles, y los cambios se producen automáticamente en segundo plano sin interrumpir su trabajo. También puede compartir blocs de notas con otras personas con unos pocos clics. Obtendrán el mismo bloc de notas, guardado en el mismo lugar.

Si prefiere utilizar AutoML para crear sus modelos, el **Piloto automático de Amazon SageMaker** crea, entrena y ajusta, automáticamente, los mejores modelos de ML con base en sus datos. También puede utilizar SageMaker JumpStart para lanzar fácil y rápidamente sus aplicaciones de ML al mercado. Con SageMaker JumpStart, puede acceder a algoritmos integrados con modelos previamente entrenados de centros modelo, FM previamente entrenados para ayudarlo a realizar tareas como el resumen de artículos y la generación de imágenes, y soluciones predesarrolladas para resolver casos de uso común. Además, puede compartir artefactos de ML, incluidos blocs de notas y modelos de ML, dentro de su organización para acelerar la creación y la implementación de modelos de ML.

Acelere el tiempo de implementación de más de 150 modelos de código abierto, incluidos los modelos de ML que se pueden implementar con un clic y los algoritmos de populares ecosistemas de modelos. Empiece ya con unos pocos clics y lance fácilmente al mercado aplicaciones de ML utilizando soluciones prediseñadas y FM preentrenados en terabytes de datos de texto e imágenes. Puede realizar una amplia gama de tareas, como el resumen de artículos y la generación de texto, imágenes o videos, que están preconfiguradas con todos los servicios de AWS necesarios para su puesta en producción, incluida una arquitectura de referencia y una plantilla de **AWS CloudFormation**.

[3] Amazon SageMaker Studio, marzo de 2023

PASO 3

Entrene modelos más rápido y a menor costo

Desafíos

Luego de que sus modelos estén creados, los equipos de ciencia de datos entrenan el modelo en sus conjuntos de datos para que estén listos para realizar predicciones precisas sobre los datos nuevos. El entrenamiento es un proceso iterativo y, con el tiempo, los modelos se deben entrenar y ajustar nuevamente para responder ante los nuevos datos o modelos. A medida que el deep learning se vuelve cada vez más generalizado, los modelos son cada vez más complejos. La complejidad de los modelos se duplica cada dos años, y muchos modelos de última generación ahora tienen billones de parámetros. El entrenamiento y el ajuste de estos modelos grandes es una tarea de computación intensiva y, a veces, los costos la imposibilita.

A medida que se amplíen los límites del rendimiento y la capacidad de los modelos de ML, el tiempo y el costo necesarios para entrenarlos seguirán en aumento. Esta fuga de recursos cada vez mayor puede impedir que su organización aproveche al máximo lo que el ML puede ofrecer, lo que frena la innovación y pone en peligro el apoyo de los ejecutivos a sus inversiones en ML.

> **Mediante el uso de instancias P4d de Amazon EC2, pudimos reducir nuestro tiempo de entrenamiento para el reconocimiento de objetos en un 40 % en comparación con las instancias de GPU de la generación anterior sin ninguna modificación de los códigos existentes".**[7]

Junya Inada, Director of Automated Driving (Recognition), TRI-AD

Solución

AWS ofrece una infraestructura de ML rentable y de alto rendimiento para el entrenamiento de ML. Elija entre la gama de Amazon EC2 de CPU, GPU e instancias basadas en aceleradores personalizadas para adaptarse a los requisitos de sus casos de uso de entrenamiento de ML. Los clientes pueden usar el **Entrenamiento de modelos de Amazon SageMaker** para aprovechar al máximo esta infraestructura sin necesidad de administrarla.

AWS Trainium: las **instancias Trn1 de Amazon EC2 Trn1**, con tecnología de AWS Trainium, ofrecen el mayor rendimiento para el entrenamiento de deep learning de modelos de NLP. Proporcionan hasta un 50 % de ahorro en costos de entrenamiento con respecto a instancias comparables de Amazon EC2. Estas instancias admiten hasta 1600 Gbps (Trn1n) de ancho de banda de red de Elastic Fabric Adapter (EFA). Se implementan en UltraClusters de Amazon EC2, con la posibilidad de escalar hasta 30 000 aceleradores AWS Trainium que están interconectados con una red no bloqueante a escala de petabits, lo que permite proporcionar hasta 6,3 exaflops de computación. Los clientes pueden utilizar las instancias de Trn1 para entrenar modelos de NLP, CV y recomendadores en muchas aplicaciones, como el resumen de textos, las recomendaciones y la generación de imágenes y videos.

GPU de NVIDIA: AWS ofrece una amplia gama de instancias basadas en GPU de NVIDIA. Las **instancias P4d Amazon EC2** son las instancias basadas en GPU más eficaces para el entrenamiento de deep learning. Son muy adecuados para el entrenamiento altamente eficaz de los modelos de ML multinodo más complejos. Las instancias P3 de Amazon EC2 son ideales cuando necesita entrenar modelos de tamaño mediano a grande y para casos de uso de entrenamiento distribuido en un solo nodo. Las **instancias G5 de Amazon EC2** brindan un costo hasta un 15 % menor de entrenamiento que las instancias P3 de Amazon EC2.

Intel/Habana Gaudi: las **instancias DL1 de Amazon EC2**, con la tecnología de los aceleradores Gaudi de Habana Labs (una empresa de Intel) están específicamente diseñadas para el entrenamiento de modelos de deep learning. Estas instancias ofrecen hasta un 40 % más de rendimiento por precio que las instancias comparables de Amazon EC2 y son muy adecuadas para casos de uso de NLP y CV.

[4] Instancias P4d Amazon EC2, marzo de 2023

Solución

Consulte la tabla de abajo para comparar las opciones de infraestructura de AWS optimizadas para el entrenamiento y el ajuste de ML.

Tipo de instancia	Cantidad máxima de chips por instancia	Tipo de acelerador	Ancho de banda de red	Almacenamiento	Características adicionales
Trn1 de Amazon EC2	16 aceleradores AWS Trainium	AWS Trainium	1600 Gbps EFA (Trn1n) 800 Gbps EFA (Trn1)	8 TB NVMe	Se puede implementar en UltraClusters de Amazon EC2 compuestos por más de 30 000 aceleradores AWS Trainium, redes de alta velocidad y almacenamiento de alto rendimiento y baja latencia Admite marcos de ML populares con el **SDK AWS Neuron**
P4d de Amazon EC2	8 GPU A-100	NVIDIA	400 Gbps EFA, GPU-Direct RDMA	8 TB NVMe	Se puede implementar en Amazon EC2 UltraClusters compuestos por más de 4000 GPU, redes de alta velocidad y almacenamiento de alto rendimiento y baja latencia
P3 de Amazon EC2	8 GPU Tesla V100	NVIDIA	100 Gbps, EFA	1,8 TB NVMe	Admite todos los marcos importantes de ML
DL1 de Amazon EC2	8 aceleradores Gaudi	Habana Labs, Intel	400 Gbps, ENA	8 TB NVMe	Admite marcos de ML populares con el SDK de Habana SynapseAI

SageMaker reduce el tiempo y el costo de entrenar y ajustar los modelos de ML mediante herramientas integradas para administrar y hacer un seguimiento de los experimentos de entrenamiento, elegir automáticamente los hiperparámetros óptimos, depurar los trabajos de entrenamiento y monitorear la utilización del ancho de banda de la red y los recursos del sistema subyacentes. SageMaker puede escalar o reducir verticalmente de forma automática la infraestructura en función de los requisitos de su trabajo de entrenamiento, de un acelerador a miles o de terabytes a petabytes de almacenamiento. Además, como solo paga por lo que utiliza, puede administrar mejor sus costos de entrenamiento.

Para entrenar los modelos de deep learning con mayor rapidez, puede utilizar el **Compilador de entrenamiento de Amazon SageMaker** para acelerar el proceso de entrenamiento de modelos hasta en un 50 % a través de optimizaciones en el nivel de gráficos y núcleos que favorecen el uso más eficiente de los aceleradores. Además, puede agregar paralelismo de datos o de modelos a su script de entrenamiento con unas pocas líneas de código, y las bibliotecas de entrenamiento distribuido de SageMaker dividirán automáticamente los conjuntos de datos de entrenamiento y los modelos en instancias de Amazon EC2 para ayudarle a completar el entrenamiento distribuido más rápidamente.

PASO 4

Implemente modelos rápidamente a un costo rentable

Desafíos

Una vez que haya entrenado y optimizado su modelo hasta el nivel deseado de exactitud y precisión, es el momento de llevarlo a producción para hacer predicciones. Esto se conoce como el paso de predicción o inferencia del ML.

Un modelo que tarda varios cientos de milisegundos en generar traducciones de texto, aplicar filtros a las imágenes o generar recomendaciones de productos puede hacer que una aplicación resulte lenta o frustrante de usar, lo que alejará a los usuarios. Al acelerar la inferencia, se puede reducir la latencia general de la aplicación y ofrecer una experiencia sin complicaciones.

Hasta el 90 % del costo de la infraestructura para desarrollar y ejecutar una aplicación de ML se invierte en la inferencia, por lo que la necesidad de una infraestructura de inferencia de ML de alto rendimiento y bajo costo es fundamental.[5]

[5] Instancias Inf1 de Amazon EC2, marzo de 2023

Solución

AWS ofrece una amplia gama de instancias de alto rendimiento, rentables y fáciles de usar para la inferencia de ML. Para modelos sumamente sofisticados, como los LLM o los modelos de difusión, las **instancias Inf2 de Amazon EC2** con tecnología de AWS Inferentia2 son la mejor opción. Las instancias Inf2 ofrecen hasta un 40 % más de rentabilidad en cuanto al precio, hasta tres veces más rendimiento y hasta ocho veces menos latencia que las instancias comparables de Amazon EC2. Las **instancias Inf1 de Amazon EC2**, con tecnología de AWS Inferentia de primera generación, son adecuados para modelos de NLP y visión más pequeños. Ofrecen hasta un 70 % menos de costo y un rendimiento 2,3 veces mayor que las instancias comparables de Amazon EC2.

Los clientes que deseen seguir utilizando el ecosistema de NVIDIA para su inferencia debido a la compatibilidad con el modelo, el marco o el operador pueden aprovechar las **instancias G5 de Amazon EC2** para una inferencia de alto rendimiento. Si busca inferencia para modelos que aprovechan las instrucciones de redes neuronales vectoriales Intel AVX-512, las **instancias C5 de Amazon EC2** pueden ayudar a acelerar las operaciones típicas de ML, como la convolución, y mejorar automáticamente el rendimiento de la inferencia en una amplia gama de cargas de trabajo de deep learning.

Utilice la tabla de abajo para comparar las opciones de infraestructura de AWS optimizadas para la inferencia de ML.

Tipo de instancia	Número máximo de aceleradores por instancia	Tipo de hardware	Ancho de banda de red	Almacenamiento	Características adicionales
Inf2 de Amazon EC2	12 aceleradores AWS Inferentia2	AWS Inferentia2	100 Gbps	40 Gbps de ancho de banda EBS	Inferencia distribuida con conectividad de alta velocidad entre aceleradores; muy conveniente para modelos ultragrandes con cientos de miles de millones de parámetros. Admite marcos de ML populares con el **SDK AWS Neuron**
Inf1 de Amazon EC2	16 aceleradores AWS Inferentia	AWS Inferentia	100 Gbps	19 Gbps de ancho de banda EBS	Admite marcos de ML populares con el **SDK AWS Neuron**
G5 de Amazon EC2	8 NVIDIA A10G Tensor Core GPUs	NVIDIA	100 Gbps	7.6 NVMe	Compatible con todos los marcos y librerías NVIDIA más importantes
C5 de Amazon EC2	96 vCPU	Intel AVX	25 Gbps	4 x 900 NVMe SSD	Creado en Nitro

Solución

SageMaker le ayuda a aprovechar la amplia selección de infraestructura de ML mencionada anteriormente y proporciona opciones de implementación de modelos para ayudarle a satisfacer sus necesidades, ya sea en tiempo real o por lotes. Una vez que implementa un modelo, SageMaker crea puntos de conexión persistentes para integrar en sus aplicaciones para hacer predicciones de ML. Admite todo el espectro de la inferencia, desde la baja latencia (unos pocos milisegundos) y el alto rendimiento (cientos de miles de solicitudes de inferencia por segundo) hasta la inferencia de larga duración para casos de uso, como el NLP. Tanto si aporta sus propios modelos y contenedores como si utiliza los proporcionados por AWS, puede implementar las prácticas recomendadas de MLOps mediante SageMaker para reducir la carga operativa de la administración de modelos de ML a escala.

Para casos de uso con patrones de uso intermitentes e impredecibles, la **inferencia sin servidor de Amazon SageMaker** le permite implementar modelos de ML con precios de pago por uso sin preocuparse por servidores o clústeres. Al implementar su modelo, con solo seleccionar la opción sin servidor, SageMaker aprovisiona, escala y desactiva de forma automática la capacidad de computación basada en el volumen de las solicitudes de inferencia, para que no necesite administrar políticas complejas de escalado ni pronosticar la demanda de tráfico por adelantado.

El **Recomendador de inferencias de Amazon SageMaker** le ayuda a elegir la mejor configuración e instancia de computación disponible para implementar modelos de ML con un costo y un rendimiento de inferencia óptimos. SageMaker Inference Recommender selecciona automáticamente el tipo de instancia de computación, la cantidad de instancias, los parámetros de los contenedores y las optimizaciones de los modelos para la inferencia a fin de maximizar el rendimiento y minimizar los costos.

Las características de implementación de modelos de SageMaker se integran de forma nativa con las capacidades de MLOps, incluidas las Canalizaciones de Amazon SageMaker (automatización y orquestación de flujos de trabajo), **Proyectos de Amazon SageMaker** (plantillas a fin de estandarizar entornos de desarrollo para científicos de datos y sistemas de integración continua y entrega continua [CI/CD] para ingenieros de MLOps), Almacén de características de Amazon SageMaker (administración de características), **Registro de modelos de Amazon SageMaker** (catálogo de modelos y artefactos para realizar un seguimiento del linaje y admitir flujos de trabajo de aprobación automatizados), **Amazon SageMaker Clarify** (detección de sesgos) y **Monitor de modelos de Amazon SageMaker** (detección de desviaciones de modelos y conceptos).

Como resultado, ya sea que implemente uno o decenas de miles de modelos, SageMaker ayuda a aliviar la sobrecarga operativa de implementar, escalar y administrar modelos de ML mientras los pone en producción más rápidamente.

> **"Lanzamos un servicio de chatbot de IA a gran escala en las instancias Inf1 de Amazon EC2 y redujimos nuestra latencia de inferencia en un 97 % con respecto a instancias comparables basadas en GPU, al tiempo que redujimos los costos. Conforme seguimos ajustando periódicamente los modelos de NLP diseñados a la medida, también es importante reducir los tiempos y costos de entrenamiento de los modelos. Con base en nuestra experiencia de migración exitosa de la carga de trabajo de inferencia en instancias Inf1 y nuestro trabajo inicial en instancias Trn1 de EC2 basadas en AWS Trainium, prevemos que las instancias Trn1 aportarán un valor adicional en la mejora del rendimiento y el costo de ML de extremo a extremo".**[6]
>
> **Takuya Nakade**, CTO, Money Forward, Inc.

[6] Instancias Trn1 de Amazon EC2, marzo de 2023

Cree una base sólida para el éxito del machine learning

La elección correcta de servicios e infraestructura puede mejorar sustancialmente el rendimiento de sus cargas de trabajo de ML, ya que podrá preparar los datos para ML más rápidamente, equiparse para crear modelos sofisticados de forma fiable, entrenar los modelos rápidamente y a escala, e implementarlos de forma potente y rentable. Tanto si quiere descargar la mayor parte del desarrollo a un servicio totalmente administrado como si quiere crear modelos desde cero, o cualquier cosa intermedia, los servicios y la infraestructura correctos pueden ayudarlo a completar los proyectos de ML más rápidamente y a lograr mejores resultados.

AWS ofrece la combinación ideal de infraestructura y servicios de alto rendimiento y bajo costo optimizados para ML. Al ejecutar sus cargas de trabajo de ML en la nube, obtendrá acceso bajo demanda a la infraestructura y las herramientas de ML que pueden crear instancias en cuestión de minutos y escalar a miles de instancias, y solo pagará por lo que utilice.

Introducción al ML ›