



# LAS 6 PRINCIPALES TENDENCIAS EN DATA SCIENCE Y ANALÍTICAS DE DATOS PARA 2023

Cómo el Data Cloud acelera el aprendizaje automático



CHAMPION  
GUIDES

EBOOK

# ÍNDICE

- 3** Introducción
- 4** Céntrese en prácticas repetibles de ML de producción
- 5** Tendencia n.º 1: herramientas e infraestructura unificadas para SQL y Python
- 6** Tendencia n.º 2: gestión e implementación de funciones de aprendizaje automático a escala con almacenes de funciones
- 7** Tendencia n.º 3: los analistas y científicos de datos obtienen el poder de los ingenieros de ML
- 8** Tendencia n.º 4: más casos de uso de datos no estructurados
- 10** Tendencia n.º 5: los científicos de datos también se convierten en desarrolladores de aplicaciones con Python
- 11** Tendencia n.º 6: crecimiento continuo de bibliotecas de ML de código abierto, herramientas y marcos
- 12** Acelere el aprendizaje automático en 2023
- 13** Acerca de Snowflake

# INTRODUCCIÓN

La data science ha evolucionado drásticamente en los últimos diez años. Se ha expandido para abarcar los campos emergentes de la inteligencia artificial (IA) y el aprendizaje automático (machine learning, ML), y ha espolcado el auge del big data (la recogida de grandes volúmenes de datos a gran velocidad). A medida que ha ido aumentando la cobertura y la escala de la data science, el procesamiento en la nube le ha seguido el ritmo proporcionando recursos casi ilimitados para adaptarse a las necesidades de los profesionales y respaldar proyectos complejos y ambiciosos.

A pesar de que la inversión en data science y ML ha aumentado, muy pocas organizaciones han experimentado en su totalidad la ventaja competitiva o el impacto empresarial derivados de disponer de analíticas avanzadas. ¿El motivo? Aplicarlos a la producción ha seguido suponiendo un obstáculo considerable.

Los ingenieros de datos, los científicos de datos y los ingenieros de ML suelen trabajar aislados, con copias independientes de los datos, lo que genera una mayor complejidad y la ausencia de soluciones colaborativas. La distribución de datos en silos (que provoca que los datos estén distribuidos por diversos sistemas y data lakes), sobrecarga a los científicos de datos con tareas improductivas que consumen mucho tiempo. Por otra parte, la complejidad del procesamiento de datos centra los esfuerzos de la organización en gestionar la infraestructura, en vez de en generar valor.

Pero el cambio ya está en marcha. Los últimos avances tecnológicos están preparados para influir significativamente en la forma de trabajar de los científicos y los analistas de datos. En 2023, 6 tendencias emergentes tienen el potencial de acelerar el ML y hacer que las organizaciones pasen de las analíticas descriptivas y diagnósticas, que explican qué ha pasado y por qué, a los análisis predictivos y prescriptivos, que pronostican qué sucederá y proporcionan valiosas indicaciones sobre cómo cambiar el futuro.

En este eBook, aprenderá que:

- La infraestructura unificada que admite varios lenguajes de programación permite a los científicos, ingenieros y analistas de datos aprovechar el máximo potencial de cada uno.
- Snowflake Data Cloud puede ampliar el acceso a los datos, el data sharing y el uso de varios tipos de datos, incluidos los no estructurados (en vista previa), a través de un ecosistema seguro con acceso a datos de terceros listos para usar.
- Los almacenes de funciones permiten a los científicos de datos gestionar e implementar funciones de ML a escala ofreciendo reproducibilidad, detectabilidad y escalabilidad.
- Los analistas y los científicos de datos son cada vez más capaces de usar el potencial de los sistemas, antes reservado a los ingenieros de ML, para facilitar y participar en procesos de producción más especializados.
- El desarrollo de aplicaciones web en Python es cada vez más accesible, lo que permite a los científicos de datos crear modelos de datos más completos y prácticos.
- Los rápidos avances en bibliotecas de código abierto, herramientas y marcos demuestran la necesidad de contar con una solución que garantice las inversiones en data science y ML en el futuro.

# CÉNTRESE EN PRÁCTICAS REPETIBLES DE ML DE PRODUCCIÓN

En los últimos años, se han realizado enormes inversiones en data science, IA y ML basadas en la promesa de aumentar la rentabilidad financiera, mejorar la eficiencia de los procesos e incrementar la resiliencia empresarial general. Según la [Encuesta Global de McKinsey sobre IA](#), la adopción de la IA ha aumentado más del doble desde 2017, y las empresas que más invierten en IA se sitúan por delante de su competencia. El valor de la data science se da por sentado: en 2023, las empresas centrarán su atención en la optimización y la eficacia de la producción.

Las empresas también se dedicarán a maximizar el efecto de las iniciativas de IA, de ML y de data science. Para ello, perfeccionarán procesos, modernizarán los sistemas heredados y adoptarán herramientas cohesivas disponibles para aprovechar lo que ofrecen estos ámbitos. Los obstáculos que durante mucho tiempo han separado la data science de los usuarios empresariales, los desarrolladores de distintos lenguajes de programación y otros profesionales están desapareciendo rápidamente.

Los avances conseguidos en 2022 apuntan a 6 interesantes tendencias para la data science y el ML en 2023. Existen tecnologías y herramientas emergentes que pueden acelerar el trabajo de los científicos de datos, eliminar los silos de datos y ampliar las posibilidades de los datos estructurados y no estructurados por igual. Es un momento muy interesante para trabajar en este ámbito.

La nube es lo que subyace a esta aceleración. Los científicos, los ingenieros y los analistas de datos se benefician de las tecnologías basadas en la nube, que proporcionan cantidades prácticamente ilimitadas de recursos de procesamiento elásticos. Además, la nube permite eliminar silos de datos consolidando data lakes, almacenes de datos y data marts para lograr que el data sharing y el análisis de datos sean sencillos, rápidos y seguros, y se realicen en una única ubicación para respaldar la colaboración entre los equipos que trabajan con datos.

**En resumen, los datos ofrecen más aplicaciones prácticas que nunca, y las nuevas herramientas se están sirviendo de esta realidad para generar cambios notables.** Por consiguiente, las organizaciones están a punto de movilizar los datos no solo para predecir el futuro, sino también para aumentar la probabilidad de ciertos resultados mediante el uso de analíticas prescriptivas.

A continuación, mostramos 6 tendencias que determinarán la data science en 2023 y que continuarán la evolución de las analíticas hacia el ML.



# TENDENCIA N.º 1: HERRAMIENTAS E INFRAESTRUCTURAS UNIFICADAS PARA SQL Y PYTHON

Conforme la cantidad de datos y aplicaciones aumentaba, los almacenes de datos on-premise y otros heredados impedían aplicar escalabilidad. La solución del sector consistió en copiar los datos en almacenes de objetos en la nube e implantar una infraestructura de procesamiento para lenguajes de programación como Python, SQL y Java. Esto complicó la gestión de la infraestructura y limitó la colaboración entre equipos.

SQL y Python son los lenguajes de programación por excelencia en el panorama de datos moderno, y cada uno brinda ventajas únicas. La velocidad de consulta y agregación de SQL es un gran beneficio, mientras que la capacidad de Python de gestionar transformaciones y analíticas complejas mediante su completo ecosistema de código abierto es una necesidad para muchas organizaciones. Cada lenguaje ofrece ventajas claras, si bien pertenecen esencialmente a mundos distintos. Además, cada uno se ejecuta en una infraestructura independiente y se desarrolla con herramientas diferentes, lo que lleva mucho tiempo impidiendo a los científicos de datos aprovechar lo mejor de ambos.

Esta ausencia de interoperabilidad ha generado un importante efecto silo, por el que los usuarios de un lenguaje no tienen la capacidad de colaborar en análisis ni flujos de trabajo con los usuarios del otro. Incluso a quienes se manejan con soltura en ambos lenguajes puede resultarles frustrante y tedioso cambiar de código y lidiar con la fricción de realizar consultas entre lenguajes.

Pese a todo, las distancias entre SQL y Python se están acortando gracias a herramientas como dbt, Hex y Snowflake Snowpark, que combinan los lenguajes para realizar cualquier tarea de datos, aprovechando así las ventajas de ambos para realizar diversas operaciones durante los análisis. Esto se consigue de la siguiente manera:

**dbt:** flujo de trabajo de transformación de datos para equipos de datos que sigue las prácticas recomendadas de ingeniería de software, como los principios de modularidad, portabilidad, integración y desarrollo continuos (continuous integration/continuous development, CI/CD) y documentación en el almacén de datos en la nube. Aunque se trata de un flujo de trabajo que da prioridad a SQL, en 2022, dbt incorporó Python como segundo lenguaje para dar respuesta a la creciente demanda de soluciones que ofrecieran fluidez al trabajar con varios lenguajes en un mismo proyecto.

**Hex:** plataforma moderna para analíticas y data science que facilita el acceso a los datos, su análisis en cuadernos colaborativos basados en Python y SQL, y compartir el trabajo como aplicaciones e historias interactivas basadas en datos. Para proporcionar una escalabilidad de procesamiento casi ilimitada, su enfoque consiste en trasladar el procesamiento al almacén de datos en lugar de cargar todos los datos en un cuaderno.

**Snowpark:** nuevo marco de desarrollo para Snowflake. Permite a los ingenieros, científicos y desarrolladores de datos escribir código en su lenguaje de preferencia y ejecutarlo en Snowflake. Este admite interfaces de desarrollo en SQL, Python y Java, entre otros, y permite cambiar de contexto sin tener que trasladar datos ni configurar clústeres independientes.

Las herramientas que unifican los lenguajes de programación son cruciales para disfrutar de una colaboración que fomente un crecimiento continuo. Estamos en 2023, y los ingenieros, científicos y analistas de datos ya no tienen que trabajar aislados debido a la falta de un lenguaje común: ahora pueden trabajar código con código para transformar los datos sin procesar en información. En última instancia, esta capacidad de compartir conocimientos fomenta que los proyectos sean más ágiles y tengan mejores resultados a largo plazo.

# TENDENCIA N.º 2: GESTIÓN E IMPLEMENTACIÓN DE FUNCIONES DE ML A ESCALA CON ALMACENES DE FUNCIONES

Cuando los científicos de datos desarrollan nuevos modelos de ML, se enfrentan a la dura tarea de preparar los datos y crear funciones. Estas se crean mediante la obtención y la preparación de columnas de datos en un formato específico que puede introducirse en los modelos de ML. Una vez que se generan las funciones de un modelo, los científicos de datos se enfrentan al reto adicional de reescribir las mismas o de dedicar tiempo a buscar y encontrar otras ya existentes para utilizarlas en el siguiente modelo.

Afortunadamente, en 2022, se produjo un acusado repunte en la adopción de almacenes de funciones (repositorios centrales que contribuyen al aumento de la capacidad de búsqueda, la colaboración y la escalabilidad de las funciones de ML). Los científicos de datos pueden encontrar con rapidez funciones transformadas y listas para usar, lo que da lugar a una experimentación más rápida y a la aceleración del proceso para la producción. Entre las ventajas de un almacén de funciones, se incluye la capacidad de aumentar la colaboración entre equipos mediante la reutilización del trabajo de otros científicos de datos. Además, los almacenes de funciones reducen el tiempo y el esfuerzo necesarios para implementar un modelo entrenado en un entorno de producción, puesto que los científicos de datos ya no necesitan redefinir lo que suele ser un flujo de datos existente.

Hoy en día, los almacenes de funciones se consideran la mejor forma de optimizar modelos de ML, puesto que los equipos pueden acceder con mayor facilidad a datos mejorados y perfeccionados que son relevantes para sus modelos. Sin embargo, crear un almacén de funciones no es una tarea fácil. Poner en práctica las funciones es todo un desafío, ya que es necesario incorporar reproducibilidad, detectabilidad y escalabilidad al almacén de funciones.

1. La **reproducibilidad de modelos** requiere que las funciones estén centralizadas en una única ubicación y que tanto estas como los datos tengan versiones. Los científicos de datos deben ser capaces de retroceder en el tiempo para descubrir qué funciones y datos se han utilizado para entrenar un modelo. Las funciones también se deben definir una vez y estar disponibles para todos los futuros casos de uso, lo que implica que los almacenes de funciones se deben actualizar con regularidad y que es necesario gestionar el control de versiones.
2. La **detectabilidad** requiere que la creación de funciones se realice fuera de las instancias de cuaderno individuales y se centralice en un repositorio unificado. Para permitir la colaboración y contribuir a la reutilización eficiente del almacén de funciones, debe existir un catálogo que permita detectar y buscar fácilmente dichas funciones.
3. La **escalabilidad** implica que los elementos del almacén de funciones pueden mantener el ritmo de cargas operativas exigentes conforme aumenten los casos de uso del ML. Un almacén de funciones debe ser capaz de escalar de cientos a millones de

funciones y seguir sirviendo con eficiencia tanto a los flujos de trabajo de entrenamiento como a los de inferencia. En lugar de ejecutar flujos duplicados de entrenamiento e inferencia, el procesamiento centralizado acelera el acceso a las funciones y reduce el procesamiento redundante.

Snowflake ofrece dos enfoques para el desarrollo de almacenes de funciones, si bien ambos evitan la creación de nuevos sistemas o silos de datos entre los científicos y los analistas de datos.

El primer enfoque consiste en aplicar una solución de código abierto, como Feast, a modo de interfaz de almacén de funciones, mientras que Snowflake se convierte en el almacén y el motor de las funciones. Las funciones persisten en la plataforma de datos única, respaldadas por cualquier herramienta existente de ingesta, catalogación y extracción, carga y transformación (extract, load, transform; ELT). Con esta solución, las organizaciones mantienen en una ubicación escalable y centralizada tanto la gestión de los flujos de datos como la interfaz en la que los científicos de datos detectan y acceden a los datos y las funciones.

Por otra parte, el segundo enfoque consiste en añadir a Snowflake una solución de almacén de funciones gestionado. Los datos de los clientes conservan su formato modelado y sin procesar en Snowflake Data Cloud y, mientras su transformación se ejecuta en Snowflake usando SQL o Snowpark para Python, el proveedor del almacén de funciones abstraer la gestión y coordinación del flujo. Tecton e Iguazio son algunos de los partners de Snowflake en este ámbito.

# TENDENCIA N.º 3: LOS ANALISTAS Y CIENTÍFICOS DE DATOS OBTIENEN EL PODER DE LOS INGENIEROS DE ML

Python es cada vez más popular en un amplio abanico de sectores. Además de los desarrolladores de ML (que tradicionalmente han recurrido a este lenguaje para gestionar una cantidad masiva de solicitudes de datos), también lo usan los científicos de datos para crear modelos y sistemas fiables basados en un código intuitivo y flexible. La popularidad de Python sigue en alza a medida que se convierte en la *lengua franca* de la data science. De hecho, el 70 % de los desarrolladores de ML y científicos de datos afirman usar Python hoy en día.

Históricamente, ejecutar Python a nivel empresarial era difícil debido a la complejidad de su infraestructura y los riesgos de seguridad de su ecosistema de código abierto. Por eso, a pesar de su escalabilidad y rendimiento, es un lenguaje que se ha reservado a los desarrolladores de ML, que no tienen problemas en usar infraestructuras complejas y lidiar con la implementación de parches de seguridad para aplicaciones esenciales. Aun así, gracias a los nuevos marcos de desarrollo que eliminan la carga de gestionar la infraestructura de Python (además de la de otros lenguajes), los científicos y analistas de datos también lo tienen más fácil que nunca para aplicar Python a escala.

Puesto que Python abarca todas las funciones de datos, los equipos de análisis de datos pueden aprovechar su velocidad y sus bibliotecas de código abierto para participar en la limpieza y la estructuración de los datos. Gracias a una mayor participación en esta área, se mitigan errores, se mejora la productividad y se obtiene mejor información para posibilitar una toma de decisiones de alta calidad.

La colaboración interdisciplinar no se limita a los equipos de análisis de datos, sino que los científicos de datos versados en Python probablemente también podrán ampliar sus funciones. En lugar de recurrir a Python para las tareas de exploración, podrán desempeñar un rol cada vez más activo en las labores relacionadas con la producción. Así, vemos cómo se tiende un puente crucial entre el mundo del desarrollo y el de la producción gracias al acceso a infraestructuras robustas que pueden abordar ambos aspectos (con una sólida base de datos que proporciona acceso específico a todos los datos, residen donde residan, y a un procesamiento

ampliable según las necesidades). Sin retrasos a la hora de crear clústeres, los científicos de datos podrán trabajar a mayor escala con información sobre cómo se trasladarán a producción sus modelos.

La incorporación de AutoML a plataformas escalables también permitirá a los científicos y analistas de datos disponer de conocimientos limitados sobre aprendizaje automático para poder entrenar el modelo y generar información importante basada en ML de forma rápida. Las herramientas de AutoML automatizan tareas asociadas al desarrollo y la implementación de modelos de ML, que tradicionalmente solo ejecutaban los expertos en data science. Permiten a más usuarios de datos automatizar una o varias partes del flujo de trabajo de ML, entre las que se incluyen la preparación de los datos, así como el entrenamiento y la selección de modelos, entre otras.

**Estas herramientas están marcando una gran diferencia para científicos y analistas de datos por igual, puesto que les permiten abordar tareas improductivas (carga, selección, preparación y limpieza de datos) que antes ocupaban el 80 % de su tiempo, pero que ahora se estima que solo ocupan el 45 %, según una encuesta realizada por Anaconda y procesada por Datanami en la que participaron científicos de datos.** Por tanto, AutoML incrementa la productividad y libera más tiempo para dedicarlo a realizar análisis.

Conforme aumente la escalabilidad y transparencia de las herramientas de AutoML, es probable que su adopción crezca, lo que permitirá a más miembros de equipos de datos aprovechar el potencial del ML en producción.



# TENDENCIA N.º 4: MÁS CASOS DE USO DE DATOS NO ESTRUCTURADOS

Según la estimación de Statista, la cantidad total de datos creados en 2022 ascendería a 97 zettabytes, una cifra que se espera que aumente rápidamente en los próximos años. Para 2025, las estimaciones de IDC recogidas por Analytics Insight también predicen que el 80 % del volumen de datos mundial serán datos no estructurados, lo que debería hacer saltar las alarmas de las organizaciones, ya que solo el 0,5 % de estos recursos se analizan hoy en día.<sup>1</sup>

El pronóstico de estos volúmenes de datos no estructurados revela que cada vez es más necesario que los científicos de datos puedan analizar datos no estructurados junto con los datos estructurados y semiestructurados (es decir, datos recogidos con algunos elementos estructurales, pero sin formatos convencionales, como registros de usuarios y actividad web en formatos como JSON).

Lamentablemente, los datos no estructurados incluyen archivos digitales que contienen datos complejos, como es el caso de texto, imágenes, vídeos, audios, archivos PDF y formatos de archivo específicos de cada sector. Puesto que se generan cantidades ingentes de datos no estructurados a partir de fuentes como documentos escritos, no procesarlos de forma eficaz es perder una oportunidad de conseguir una ventaja competitiva. Es la complejidad de estas fuentes de datos no estructurados la que hace que analizarlos junto con otros tipos de datos sea extremadamente difícil. Sin una fuente única que admita todos los tipos de datos, los no estructurados quedan atascados en silos. Por tanto, los científicos de datos no pueden buscar, analizar o consultar fácilmente datos no estructurados, y deben obtenerlos de múltiples sistemas.

Además de los problemas de gestión de datos, resulta prácticamente imposible para cualquier organización producir o recopilar todos los datos necesarios para descubrir tendencias empresariales y competitivas. La capacidad de compartir y unir conjuntos de datos, tanto dentro de las organizaciones como entre ellas, se considera cada vez más la mejor forma de obtener un mayor valor de los datos. Este es el motivo por el que los científicos y los analistas de datos están continuamente buscando más datos para complementar sus modelos y análisis de ML con datos externos a fin de mejorar la precisión de las predicciones.



Con Snowflake, los científicos y los analistas de datos tienen acceso a un sistema global y unificado para gestionar todos los tipos de datos, incluidos los datos no estructurados. Conforme se sigan generando más y más datos no estructurados, Snowflake Data Cloud continuará proporcionando una fuente única y consolidada de datos que permita a científicos y analistas acceder a los datos y procesarlos más rápido para extraer valor de todos ellos.

Además de esta capacidad para usar varios tipos de datos, Snowflake también posibilita un acceso seguro, gobernado y fluido a datos externos, de modo que permite a quienes no usen Snowflake cumplir las obligaciones de conformidad de diversos panoramas normativos de protección de datos.

**Snowflake Data Cloud** es un ecosistema donde los clientes, los partners y los proveedores de datos y de servicios de datos de Snowflake se conectan a sus propios datos, y comparten y consumen de forma fluida datos y aplicaciones de otros usuarios. El Data Cloud cuenta con el respaldo de la plataforma de Snowflake, elimina las barreras que imponen los datos en silos y permite a las organizaciones unificar y conectarse a una única copia de los datos. Además, el Data Cloud es una forma perfecta de obtener valor de conjuntos de datos comercializados de rápido crecimiento con un acceso rápido, sencillo y controlado.

La **tecnología Snowflake Secure Data Sharing** respalda el Data Cloud, lo que permite a las empresas colaborar fácilmente con sus partners empresariales externos e internos sin los tradicionales obstáculos a las transferencias de datos. Con Snowflake, los datos nunca se copian ni se trasladan. En su lugar, el

propietario de los datos otorga a los usuarios acceso a los datos actualizados en su ubicación original. Gracias a la separación del almacenamiento y el procesamiento que ofrece Snowflake, la latencia y la contención de usuarios simultáneos nunca es un problema. Dado que los cambios en los datos se realizan en una única versión, estos están siempre actualizados para todos los usuarios y se garantiza que los modelos de datos siempre utilicen la última versión.

En Snowflake Marketplace, los usuarios pueden acceder a aplicaciones y datos de terceros actualizados y listos para consultar, así como adquirirlos. Al aplicar la misma tecnología de data sharing seguro, no se necesitan procesos de ETL, lo que permite a los científicos y analistas de datos empezar a usar datos de terceros proveedores más rápido. Snowflake Marketplace simplifica y acelera la detección y evaluación de datos de terceros con muestras de datos prácticas que se pueden combinar sin problemas con datos propios para validar prototipos. Una vez que se ha evaluado un conjunto de datos, el proceso de compra se puede realizar directamente en el producto, y los cargos adicionales se incorporan a su factura de Snowflake.

Los datos externos están disponibles y accesibles para todos los usuarios del Data Cloud con tan solo unos clics. Una vez en el Data Cloud, los datos están listos para compartirse y utilizarse. No es necesario enviar archivos CSV ni ocuparse del control manual de las versiones. Los científicos de datos pueden mejorar los modelos con un acceso ininterrumpido a datos casi ilimitados sobre cualquier tema, incluso en tiempo real y en circunstancias cambiantes.

# TENDENCIA N.º 5: LOS CIENTÍFICOS DE DATOS TAMBIÉN SE CONVIERTEN EN DESARROLLADORES DE APLICACIONES CON PYTHON

En cualquier organización, los científicos de datos son responsables de compartir datos y modelos de formas comprensibles e interesantes para sus colaboradores. Pese a ello, los científicos de datos se topan a menudo con la dificultad de mostrar datos en paneles que no admiten nuevas formas de comunicar esta información. Además, los equipos de data science no suelen disponer de desarrolladores de aplicaciones de pila completa para crear productos internos mediante los que compartir su trabajo.

En el próximo año veremos un cambio significativo del statu quo a medida que los científicos de datos adquieran capacidades de desarrollo de aplicaciones en Python, gracias a marcos de desarrollo de código abierto como Streamlit. Estos facilitan el proceso desde la concepción de la idea hasta la creación de la aplicación usando únicamente Python, lo que salva la brecha entre el ML y la acción de forma eficaz.

Los científicos de datos pueden crear rápidamente aplicaciones interactivas con componentes enriquecidos, como gráficos y campos de entrada, para permitir a sus equipos interactuar con datos y modelos sin la tradicional complejidad derivada de crear una aplicación web, como definir rutas, gestionar solicitudes HTTP y escribir código HTML, CSS o JavaScript. Esta plataforma y otras similares permiten a los científicos de datos desarrollar aplicaciones web que den vida a los modelos de ML de formas a las que los equipos de data science nunca habían accedido.

Gracias a la capacidad de crear aplicaciones rápidamente solo con Python, y compartirlas y crear iteraciones en esas plataformas interactivas, los equipos pueden movilizar los datos más fácilmente y poner la información basada en ML en manos de los usuarios empresariales para que tomen las medidas pertinentes, que es el objetivo final.

Históricamente, un abismo separaba el trabajo de los científicos de datos y los usuarios empresariales. Muchos científicos de datos dedican una cantidad de tiempo y energía considerable a desarrollar modelos cuya comprensión y aplicación práctica por parte de los usuarios empresariales no están aseguradas. Sin datos y análisis accesibles y fiables, las acciones se estancan, y la promesa de la data science nunca se materializa totalmente.

Streamlit está cambiando todo eso. Esta solución, que ya han adoptado muchas empresas de la lista Fortune 500, democratiza el desarrollo de aplicaciones web de forma eficaz. Puesto que Snowflake ha adquirido Streamlit, las preocupaciones sobre la gestión de la infraestructura, la elasticidad y la gobernanza de datos de seguridad de estas aplicaciones son cosa del pasado. Como parte de la iniciativa de integración continua, los usuarios pueden implementar con fluidez y compartir de forma segura sus aplicaciones aprovechando la seguridad y fiabilidad de la infraestructura de Snowflake.

Conforme más organizaciones vayan adoptando plataformas como Streamlit para capacitar a sus científicos de datos como desarrolladores de aplicaciones, se acelerará el desarrollo de modelos, mejorará la colaboración y se compartirá información de forma más eficiente. Esto proporciona agilidad y capacidad de acción a las organizaciones, lo que respalda el crecimiento a largo plazo.

# TENDENCIA N.º 6: CRECIMIENTO CONTINUO DE BIBLIOTECAS DE ML DE CÓDIGO ABIERTO, HERRAMIENTAS Y MARCOS

El campo de la data science está evolucionando a pasos agigantados. No solo se publican nuevos desarrollos de ML e IA cada mes (muchos de los cuales son de código abierto), sino que también surgen regularmente nuevas empresas emergentes, herramientas y soluciones. Un ejemplo de ello es el rápido auge y la sorprendentemente veloz adopción de ChatGPT. El bot de chat desarrollado por OpenAI con un modelo lingüístico optimizado se ha convertido rápidamente en un fenómeno viral que ha incorporado la IA de código abierto al ámbito público por primera vez. Con el rápido ritmo de innovación, adopción y cambio que está siguiendo este campo, es imperativo no quedarse anclado en el uso de una única herramienta.

Por eso, también es importante seleccionar una plataforma independiente de algoritmos y marcos que pueda interactuar de forma cohesiva con otras herramientas. Al elegir una plataforma preparada para el futuro, se asegura de que las próximas herramientas de ML sigan funcionando sin problemas en su plataforma actual. Al fin y al cabo, lo último que se quiere es tener que cambiar de plataforma solo para usar la siguiente generación de herramientas.

Lo que hace única a la plataforma de Snowflake es su arquitectura moderna. Diseñada con un procesamiento y un almacenamiento independientes, pero integrados de forma lógica, Snowflake acaba con los esfuerzos de creación manual de clústeres que otros sistemas

deben llevar a cabo para lograr que las distintas capas funcionen juntas. Como resultado, Snowflake ofrece una **arquitectura de datos compartidos y multiclúster** que proporciona una escalabilidad casi ilimitada, elasticidad instantánea y niveles sumamente elevados de simultaneidad para respaldar todo el procesamiento que necesitan la data science y el ML, desde la preparación de datos hasta la inferencia y la implementación de modelos en una aplicación.

Además de la arquitectura subyacente que respalda todos los tipos de datos y reúne varios lenguajes para el procesamiento en un solo motor, Snowflake admite integraciones con el ecosistema de data science de diversas maneras.

- La función **External Functions** de Snowflake permite a los usuarios interactuar con cualquier servicio de ML personalizado, alojado o de terceros externo a Snowflake usando SQL. Este podría ser un servicio de IA de conversión de voz a texto u otros servicios de procesamiento del lenguaje natural (PLN), o incluso un modelo de ML que se implemente en un entorno externo para solicitudes de predicción en tiempo real.
- Para desarrollar e implementar el código de Python y Java con **Snowpark**, los desarrolladores siempre han tenido la flexibilidad de trabajar con su entorno de desarrollo integrado (integrated development environment, IDE) o cuaderno favorito. Al usar la biblioteca cliente de Snowpark, los desarrolladores pueden trasladar el procesamiento de casi cualquier herramienta de desarrollo a Snowflake, de forma que no es necesario dejar de utilizar las herramientas preferidas.

- Puesto que la potencia de Python reside en su completo ecosistema de paquetes de código abierto, como parte de la oferta de Snowpark para Python, nos complace incorporar innovaciones óptimas de código abierto y nivel empresarial al Data Cloud mediante la integración con Anaconda. Gracias al completo conjunto de paquetes de código abierto y la gestión fluida de dependencias de Anaconda, podrá acelerar los flujos de trabajo basados en Python.

Además, Snowflake también ofrece integraciones con populares herramientas de código abierto en las numerosas capas de la pila de aprendizaje automático. Esto incluye integraciones con las conocidas tablas de Apache Iceberg (en vista previa) para acceder a todos sus datos más fácilmente a fin de disponer de soluciones de almacén como Feast, que ayudan a gestionar el ciclo de vida integral de las funciones de ML.

- Con el crecimiento de los datos no estructurados, llega también el desarrollo paralelo de métodos para gestionarlos y procesarlos. Por ejemplo, los nuevos servicios de etiquetado permiten etiquetar imágenes y otros datos no estructurados manualmente. Con **Snowflake Secure Data Sharing**, los datos no estructurados (en vista previa pública) se pueden compartir con un proveedor para añadirles etiquetas sin tener que moverlos. Además, se pueden emplear herramientas de análisis de datos no estructurados sobre Snowflake para mejorar dichos datos usando los servicios de PLN ofrecidos por empresas como Hugging Face y los servicios de IA en la nube como, por ejemplo, AWS Rekognition.

# ACELERE EL APRENDIZAJE AUTOMÁTICO EN 2023

Es sorprendente la rapidez con la que la data science se ha convertido en algo tan común. En los últimos diez años, las empresas han ampliado sus miras y, aparte de centrarse en la elaboración de informes y el análisis histórico, han integrado la data science con ML y modelos matemáticos avanzados. No obstante, la mayoría de las organizaciones que usan el ML todavía no han obtenido retorno de la inversión, pues les cuesta llevar proyectos a producción.

Una plataforma de datos es una base necesaria para proporcionar acceso a los datos y unas capacidades de procesamiento de estos a los que se pueda aplicar escalabilidad fácilmente, y que se adapten a las necesidades de seguridad más exigentes de, incluso, los sectores más regulados, como los servicios financieros y la asistencia sanitaria. Snowflake proporciona una arquitectura compatible con muchos lenguajes de programación, como SQL y Python, que posibilita la consolidación de datos, la preparación eficiente de datos, el acceso sencillo a bibliotecas de código abierto y amplias integraciones con el ecosistema de data science. Sus datos se movilizan, lo que le permite beneficiarse inmediatamente de nuevas tendencias en data science y ML.

**Con Snowflake, desaparecen la complejidad de incorporar la data science y el ML a la producción. ¿Está preparado para acelerar su ML y aprovechar todo su valor?**





## ACERCA DE SNOWFLAKE

Snowflake permite a cualquier organización movilizar sus datos con Snowflake Data Cloud. Los clientes utilizan el Data Cloud para unificar, descubrir y compartir datos de forma segura, y ejecutar diversos workloads analíticos. Independientemente de la ubicación de los datos o de los usuarios, Snowflake ofrece una experiencia de datos única que abarca varias nubes y regiones geográficas. Miles de clientes de numerosos sectores, incluidas 573 de las empresas que figuran en Forbes Global 2000 (G2K) (2022), a fecha de 31 de enero de 2023, utilizan Snowflake Data Cloud para impulsar sus negocios. Más información en [snowflake.com](https://www.snowflake.com)



© 2023 Snowflake Inc. Todos los derechos reservados. Snowflake, el logotipo de Snowflake y el resto de nombres de productos, funciones y servicios de Snowflake mencionados en este documento son marcas registradas o marcas comerciales de Snowflake Inc. en Estados Unidos y otros países. El resto de logotipos o nombres de marcas mencionados o utilizados en este documento se usan únicamente con fines identificativos, y pueden ser las marcas comerciales de sus respectivos titulares. Snowflake puede no estar asociado con, patrocinado o apoyado por cualquiera de dichos titulares.

---

### CITAS

<sup>1</sup> [bit.ly/3lkt1qx](https://bit.ly/3lkt1qx)