



DIE 6 WICHTIGSTEN TRENDS FÜR DATA SCIENCE UND ANALYTICS IN 2023

Wie die Data Cloud maschinelles Lernen beschleunigt



CHAMPION
GUIDES

EBOOK

INHALTSVERZEICHNIS

- 3** Einführung
- 4** Schwerpunkt auf wiederholbaren Praktiken für maschinelles Lernen in der Produktionsphase
- 5** Trend 1: Einheitliche Tools und Infrastruktur für SQL und Python
- 6** Trend 2: Verwalten und Bereitstellen umfangreicher ML-Features mit Feature Stores
- 7** Trend 3: Analysten und Data Scientists erhalten die Leistung von ML Engineers
- 8** Trend 4: Weitere Anwendungsfälle für unstrukturierte Daten
- 10** Trend 5: Mit Python werden auch Data Scientists zu App-Entwicklern
- 11** Trend 6: Kontinuierliches Wachstum bei Open-Source-ML-Bibliotheken, -Tools und -Frameworks
- 12** Beschleunigen des maschinellen Lernens in 2023
- 13** Über Snowflake

EINFÜHRUNG

Data Science hat sich in den letzten zehn Jahren bedeutend weiterentwickelt. Die Disziplin umfasst inzwischen auch die aufstrebenden Bereiche der künstlichen Intelligenz (KI) und des maschinellen Lernens (ML) und hat den Erfolg von Big Data – also von großen Datenmengen, die mit hoher Geschwindigkeit entstehen – vorangetrieben. Cloud-Computing hat mit der zunehmenden Reichweite und Skalierung von Data Science Schritt gehalten und erfüllt die Bedürfnisse von Experten durch die Bereitstellung nahezu unbegrenzter Ressourcen zur Unterstützung ehrgeiziger und komplexer Projekte.

Trotz steigender Investitionen in Data Science und maschinelles Lernen (ML) haben nur sehr wenige Unternehmen die volle Auswirkung ihrer fortschrittlichen Analysen auf ihr Geschäft oder ihren Wettbewerbsvorteil erkannt. Was sind die Gründe? Der Übergang in die Produktionsphase stellt nach wie vor eine große Hürde dar.

Data Engineers, Data Scientists und ML Engineers arbeiten oft isoliert mit unterschiedlichen Datenkopien. Das führt zu größerer Komplexität und einem Mangel an kollaborativen Lösungen. Datensilos, in denen die Daten über verschiedene Systeme und Data Lakes verstreut sind, belasten die Data Scientists mit zeitraubender Arbeit, während die Komplexität der Verarbeitung dazu führt, dass sich Unternehmen auf die Verwaltung der Infrastruktur konzentrieren, anstatt einen Mehrwert zu schaffen.

Doch das soll sich nun ändern. Die jüngsten technologischen Entwicklungen werden die Art und Weise, wie Data Scientists und Datenanalysten arbeiten, erheblich beeinflussen. 2023 haben sechs Trends das Potenzial, den Ausbau von ML zu beschleunigen und Unternehmen weg von deskriptiven und diagnostischen Analysen – die erklären, was passiert ist und warum – hin zu prädiktiven und präskriptiven Analysen zu führen, die vorhersagen, was passieren wird, und aussagekräftige Empfehlungen geben, wie die Zukunft verändert werden kann.

In diesem E-Book erfahren Sie:

- wie eine einheitliche Infrastruktur, die mehrere Programmiersprachen unterstützt, es Data Scientists, Data Engineers und Datenanalysten ermöglicht, das maximale Potenzial der einzelnen Sprachen zu nutzen.
- wie die Data Cloud von Snowflake den Datenzugriff, das Data Sharing und die Verwendung verschiedener Datentypen, einschließlich unstrukturierter Daten (in Preview), über ein sicheres Ökosystem mit direktem Zugriff auf Daten von Drittanbietern erweitert.
- wie Data Scientists mithilfe von Feature Stores ML-Features in großem Umfang verwalten und Bereitstellung betreiben können, indem sie für Reproduzierbarkeit, Auffindbarkeit und Skalierbarkeit sorgen.
- wie Datenanalysten und Data Scientists zunehmend in der Lage sind, die Leistungsfähigkeit von Systemen zu nutzen, die bisher ML Engineers vorbehalten waren, um intelligentere Produktionsprozesse zu fördern und daran teilzuhaben.
- wie die Entwicklung von Web-Apps in Python, die immer mehr Menschen zugänglich wird, Data Scientists die Möglichkeit gibt, ihre Modelle verständlicher und besser nutzbar zu machen.
- wie die rasanten Fortschritte bei Open-Source-Bibliotheken, -Tools und -Frameworks die Notwendigkeit einer Lösung belegen, die Data Science- und ML-Investitionen zukunftssicher machen.

SCHWERPUNKT AUF WIEDERHOLBAREN PRAKTIKEN FÜR MASCHINELLES LERNEN IN DER PRODUKTIONSPHASE

Angetrieben durch das Versprechen höherer finanzieller Erträge, effizienterer Prozesse und größerer Resilienz von Unternehmen wurden in den letzten Jahren enorme Investitionen in Data Science, KI und ML getätigt. Die Akzeptanz von KI hat sich seit 2017 mehr als verdoppelt, so der Global Survey on AI von McKinsey. Unternehmen, die mehr in KI investieren, haben einen Vorsprung vor ihren Mitbewerbern. Der Wert von Data Science ist unbestritten; in 2023 werden Unternehmen ihr Augenmerk auf eine rationalisierte Produktion und Optimierung richten.

Unternehmen werden sich auch darauf konzentrieren, die Wirkung von Maßnahmen in den Bereichen KI, ML und Data Science zu maximieren, indem sie Prozesse verfeinern, Altsysteme modernisieren und kohärente Tools zur Nutzung dieser Bereiche einsetzen. Barrieren, die lange Zeit zwischen Data Science- und geschäftlichen Benutzern, Entwicklern verschiedener Programmiersprachen und mehr bestanden, werden schnell abgebaut.

Die 2022 erzielten Fortschritte weisen auf sechs spannende Trends für Data Science und ML in 2023 hin. Neue Tools und Technologien können die Arbeit von Data Scientists beschleunigen, Datensilos beseitigen und die Möglichkeiten von strukturierten und unstrukturierten Daten gleichermaßen erweitern. Dies ist eine aufregende Zeit, um in diesem Metier zu arbeiten.

Die Cloud bildet die Grundlage für diese Beschleunigung. Data Scientists, Data Engineers und Datenanalysten profitieren von Cloud-Technologien, die elastische und praktisch unbegrenzte Mengen an Rechenleistung bereitstellen. Zudem gestattet die Cloud die Beseitigung von Datensilos durch die Konsolidierung von Data Lakes, Data Warehouses und Data Marts für das schnelle, sichere und einfache Data Sharing und die Analyse an einem einzigen Ort, um die Collaboration zwischen Datenteams zu stärken.

Kurz gesagt: Daten sind besser verwertbar als je zuvor, und neue Tools nutzen diese neue Realität, um die Entwicklung voranzutreiben. Unternehmen stehen folglich kurz davor, Daten nicht nur für Prognosen zu verwenden, sondern auch die Wahrscheinlichkeit bestimmter Ergebnisse durch präskriptive Analysen zu erhöhen.

Hier stellen wir Ihnen sechs Trends vor, die 2023 die Data Science prägen und die Entwicklung der Analytik in Richtung ML fortsetzen werden.



TREND 1: EINHEITLICHE TOOLS UND INFRASTRUKTUR FÜR SQL UND PYTHON

Als die Datenmengen und die Zahl der Anwendungen wuchsen, behinderten on-premise und andere ältere Data Warehouses die notwendige Skalierbarkeit. Die Branche löste dieses Problem, indem sie Daten in Cloud Object Stores kopierte und eine Verarbeitungsinfrastruktur für Programmiersprachen wie Python, SQL und Java einrichtete. Dies führte zu einer komplexen Infrastrukturverwaltung und eingeschränkter Collaboration zwischen den Teams.

SQL und Python sind die wichtigsten Sprachen in der modernen Datenlandschaft, und jede von ihnen bietet einzigartige Vorteile. Die Geschwindigkeit von SQL bei Abfragen und Aggregationen ist ein enormer Vorteil, während die Fähigkeit von Python, komplexe Analysen und Transformationen mithilfe seines reichhaltigen Open-Source-Ökosystems zu verwalten, für viele Unternehmen eine Notwendigkeit ist.

Beide haben klare Vorteile, dennoch befinden sie sich im Grunde in unterschiedlichen Welten. Beide laufen auf einer separaten Infrastruktur und werden mit unterschiedlichen Tools entwickelt, was Data Scientists lange daran gehindert hat, das Beste aus beiden Welten herauszuholen.

Dieser Mangel an Interoperabilität hat dazu geführt, dass Benutzer einer Sprache nicht in der Lage sind, mit Benutzern der anderen Sprache bei Analysen oder Workflows zusammenzuarbeiten. Selbst für diejenigen, die beide Sprachen sicher beherrschen, kann das Umschalten zwischen den Codes und die Schwierigkeiten bei Abfragen in beiden Sprachen frustrierend und zeitraubend sein.

Die Trennung zwischen SQL und Python wird jedoch dank Tools wie dbt, Hex und Snowflake Snowpark, die die Sprachen bei jeder Datenaufgabe kombinieren und die Vorteile beider Sprachen für verschiedene Operationen während der Analyse nutzen, immer kleiner. Und das funktioniert so:

dbt: Datentransformations-Workflow für Datenteams, die bei der Softwareentwicklung bewährte Methoden wie Modularität, Portabilität, CI/CD und Dokumentation in ihrem Cloud-Warehouse befolgen. Als Transformations-Workflow, der zunächst auf SQL basierte, führte dbt 2022 Python als zweite Sprache ein, um den wachsenden Bedarf an nahtlosen Lösungen für die Arbeit mit verschiedenen Sprachen im selben Projekt zu erfüllen.

Hex: Hex ist eine moderne Plattform für Analytik und Data Science, die es einfach macht, eine Verbindung zu Daten herzustellen, diese dann in kollaborativen Notebooks auf der Basis von SQL und Python zu analysieren und die Arbeit in Form von interaktiven Daten-Apps und Stories mit anderen auszutauschen. Um eine nahezu unbegrenzte Skalierbarkeit der Verarbeitung zu gewährleisten, werden nicht alle Daten in ein Notebook geladen, sondern die Datenverarbeitung in das Warehouse verlagert.

Snowpark: Snowpark ist ein neues Entwickler-Framework für Snowflake. Es hilft Data Engineers, Data Scientists und Datenentwicklern, Code in ihrer bevorzugten Sprache zu schreiben und diesen Code in Snowflake auszuführen. Snowflake unterstützt Schnittstellen für die Entwicklung in SQL, Python, Java und mehr und erlaubt einen einfachen Kontextwechsel, ohne dass Daten verschoben oder separate Cluster eingerichtet werden müssen.

Tools, die Programmiersprachen vereinheitlichen, sind entscheidend für kontinuierliches Wachstum durch Collaboration. In 2023 müssen Data Engineers, Data Scientists und Analysten nicht mehr isoliert arbeiten, weil ihnen eine gemeinsame Sprache fehlt. Sie können zusammenarbeiten, um von Rohdaten zu Erkenntnissen zu gelangen. Dieser Wissensaustausch führt letztendlich zu agileren Projekten mit besseren Langzeitergebnissen.

TREND 2: VERWALTEN UND BEREITSTELLEN UMFANGREICHER ML-FEATURES MIT FEATURE STORES

Data Scientists stehen bei der Entwicklung neuer ML-Modelle vor der mühsamen Aufgabe, Daten aufzubereiten und Features zu erstellen. Features werden durch Beschaffung und Aufbereitung von Datenspalten in einem bestimmten Format erstellt, die in Modelle für maschinelles Lernen eingespeist werden können. Sobald die Features für ein Modell erstellt wurden, stehen Data Scientists vor der nächsten Herausforderung, entweder dieselben Features neu zu schreiben oder Zeit damit zu verbringen, vorhandene Features zu suchen und zu finden, um sie für das nächste Modell zu verwenden.

Glücklicherweise gab es 2022 einen starken Anstieg bei der Einführung von Feature Stores – einem zentralen Repository, das die Durchsuchbarkeit, Collaboration und Skalierbarkeit von ML-Features verbessert. Data Scientists können schnell Features finden, die transformiert und einsatzbereit sind, was sowohl zu schnelleren Experimenten als auch zu kürzeren Produktionszeiten führt. Zu den Vorteilen eines Feature Stores gehört die Möglichkeit, die Collaboration zwischen Teams durch die Wiederverwendung der Arbeit anderer Data Scientists zu verbessern. Feature Stores reduzieren außerdem den Zeit- und Arbeitsaufwand für die Bereitstellung eines trainierten Modells in einer Produktionsumgebung, da Data Scientists die oft vorhandene Daten-Pipeline nicht mehr neu definieren müssen.

Heute werden Feature Stores als der beste Weg zur Verbesserung von ML-Modellen angesehen, da Teams leichter auf erweiterte und optimierte Daten zugreifen können, die für ihre Modelle relevant sind. Der Aufbau eines Feature Store ist jedoch mit etwas Aufwand verbunden. Die Operationalisierung von Features ist eine Herausforderung, da der Feature Store Reproduzierbarkeit, Auffindbarkeit und Skalierbarkeit bieten muss.

- 1. Reproduzierbarkeit von Modellen:** Features müssen an einem zentralen Ort verfügbar sein und Daten sowie Features müssen versioniert werden. Data Scientists müssen die Möglichkeit haben, zeitlich zurückzugehen, um die Features und Daten zu ermitteln, mit denen das Modell trainiert wurde. Features müssen auch einmal definiert werden und anschließend für alle künftigen Anwendungsfälle verfügbar sein. Feature Stores müssen also regelmäßig aktualisiert und die jeweiligen Versionen kontrolliert werden.
- 2. Auffindbarkeit:** Features dürfen nicht mehr auf individuellen Notebook-Instanzen erstellt werden, sondern müssen zentral in einem vereinheitlichten Repository angelegt werden. Um die Zusammenarbeit und eine effiziente Wiederverwendbarkeit im Feature Store zu begünstigen, muss ein Katalog vorhanden sein, über den Features schnell auffindbar sind.
- 3. Skalierbarkeit:** Elemente im Feature Store können ohne großen operativen Aufwand mit dem Wachstum der ML-Anwendungsfälle Schritt halten. Ein Feature Store muss von Hunderten auf Millionen Features skaliert werden können und dabei weiterhin

sowohl Trainings- als auch Inferenz-Workflows effizient unterstützen. Anstatt doppelte Pipelines für Training und Inferenz laufen zu lassen, beschleunigt die zentralisierte Verarbeitung den Zugriff auf Features und reduziert die redundante Verarbeitung.

Snowflake bietet zwei Ansätze für den Aufbau von Feature Stores. Bei beiden wird vermieden, dass neue Systeme oder Datensilos zwischen Data Scientists und Datenanalysten geschaffen werden.

Der erste Ansatz besteht darin, eine Open-Source-Lösung wie Feast als Schnittstelle für den Feature Store zu nutzen, während Snowflake als Store und Engine für die Features dient. Die Features verbleiben auf der zentralen Datenplattform und werden von allen vorhandenen Erfassungs-, ELT- und Katalogisierungstools unterstützt. Bei dieser Lösung verfügen Unternehmen sowohl über die Verwaltung der Daten-Pipelines als auch über die Schnittstelle, über die Data Scientists Daten und Features an einem zentralen, skalierbaren Ort auffinden und abrufen können.

Alternativ besteht der zweite Ansatz darin, zusätzlich zu Snowflake eine Managed Feature Store-Lösung zu nutzen. Die Kundendaten verbleiben in ihrer rohen und modellierten Form in der Snowflake Data Cloud. Während die Transformation der Daten in Snowflake mit SQL oder Snowpark for Python ausgeführt wird, erfolgen die Orchestrierung und die Verwaltung der Pipeline abstrahiert durch den Anbieter des Feature Stores. Zu den Partnern von Snowflake in diesem Bereich gehören Tecton und Iguazio.

TREND 3: ANALYSTEN UND DATA SCIENTISTS ERHALTEN DIE LEISTUNG VON ML ENGINEERS

Python findet in zahlreichen Branchen immer mehr Zuspruch und wird nicht nur von ML-Entwicklern verwendet – die traditionell auf Python setzen, weil es eine große Menge an Datenanfragen verarbeiten kann – sondern auch von Data Scientists, um vertrauenswürdige Modelle und Systeme zu entwickeln, die auf intuitivem, flexiblem Code basieren. Die Popularität von Python nimmt weiter zu, da es sich zur *Lingua Franca* der Data Science etabliert. Tatsächlich geben 70 % der ML-Entwickler und Data Scientists an, Python zu verwenden.

In der Vergangenheit war es schwierig, Python auf Unternehmensebene einzusetzen. Gründe dafür waren die komplexe Infrastruktur und die Sicherheitsrisiken des Open-Source-Ökosystems. Daher blieb die Sprache trotz ihrer Skalierbarkeit und Leistung ML-Entwicklern vorbehalten, die sich mit der Verwaltung komplexer Infrastrukturen und mit Sicherheitspatches für unternehmenskritische Anwendungen auskennen. Trotzdem machen es neue Entwickler-Frameworks, die den Aufwand für die Verwaltung der Infrastruktur für Python (neben anderen Sprachen) verringern, für Data Scientists und Analysten einfacher denn je, Python auch im großen Maßstab zu nutzen.

Mit der zunehmenden Verbreitung von Python bei Datenfunktionen können Datenanalyseteams die Geschwindigkeit und die Open-Source-Bibliotheken nutzen, um bei der Bereinigung und Strukturierung von Daten mitzuwirken. Durch eine stärkere Beteiligung in diesem Bereich werden Fehler vermieden, die Produktivität erhöht und bessere Erkenntnisse gewonnen, die eine qualitativ hochwertige Entscheidungsfindung begünstigen.

Die übergreifende Collaboration endet jedoch nicht bei den Datenanalyseteams. Data Scientists, die Python beherrschen, werden ihre Rolle wahrscheinlich ebenfalls erweitern. Anstatt sich bei Untersuchungen auf Python zu stützen, werden sie bei der produktionsbezogenen Arbeit eine immer aktivere Rolle übernehmen. Hier sehen wir eine entscheidende Brücke zwischen der Welt der Entwicklung und der Produktion dank des Zugangs zu robusten Infrastrukturen, die beides können – mit einer starken Datengrundlage, die Ad-hoc-Zugang zu sich überall

befindlichen Daten bietet, und einer Rechenleistung, die bei Bedarf leicht erweitert werden kann. Da es beim Anlegen von Clustern keine Verzögerung gibt, können Data Scientists in größerem Maßstab arbeiten und erhalten Erkenntnisse darüber, wie ihre Modelle in die Produktion übernommen werden.

AutoML, das in skalierbare Plattformen eingebettet ist, versetzt auch Data Scientists und Analysten mit begrenzten Kenntnissen auf dem Gebiet des maschinellen Lernens in die Lage, schnell ML-Projekte zu trainieren und implementieren, um aussagekräftige Erkenntnisse zu erhalten. AutoML besteht aus Tools, die Aufgaben im Zusammenhang mit der Entwicklung und der Bereitstellung von ML-Modellen automatisieren, die traditionell nur von erfahrenen Data Scientists ausgeführt werden. Dadurch können immer mehr Datennutzer einen oder mehrere Teile des ML-Workflows automatisieren, z. B. die Datenaufbereitung, das Modelltraining und die Modellauswahl und vieles mehr.

Diese Tools verändern die Arbeit von Data Scientists und Datenanalysten grundlegend, denn sie nehmen ihnen die Arbeit ab (Laden, Auswählen, Aufbereiten und Bereinigen von Daten), die früher 80 % ihrer Zeit in Anspruch genommen hat. Laut einer Befragung von Data Scientists, die von Anaconda durchgeführt wurde und über die Datanami berichtet, sind es jetzt nur noch 45 %. So steigert AutoML die Produktivität und bietet mehr Zeit für die Durchführung von Analysen.

Da AutoML-Tools immer skalierbarer und transparenter werden, wird ihre Verbreitung wahrscheinlich zunehmen, sodass mehr Beteiligte im Datenteam die Möglichkeiten von ML in der Produktion nutzen können.



TREND 4: WEITERE ANWENDUNGSFÄLLE FÜR UNSTRUKTURIERTE DATEN

Statista schätzt die insgesamt in 2022 erzeugte Datenmenge auf 97 Zettabytes, eine Zahl, die in den kommenden Jahren rapide ansteigen dürfte. Laut IDC-Prognosen, die von Analytics Insight veröffentlicht wurden, werden bis 2025 80 % der weltweiten Daten unstrukturiert sein. Bei den Unternehmen sollten deswegen die Alarmglocken läuten, da heute nur 0,5 % dieser Ressourcen analysiert werden.¹

Diese zu erwartenden Mengen an unstrukturierten Daten zeigen, dass Data Scientists zunehmend in der Lage sein müssen, neben strukturierten und halbstrukturierten Daten auch unstrukturierte Daten zu analysieren (d. h. erfasste Daten, die zwar einige Strukturelemente enthalten, aber nicht auf herkömmliche Weise formatiert sind, wie z. B. Benutzerprotokolle und Webaktivitäten, die in Formaten wie JSON geliefert werden).

Leider gehören zu den unstrukturierten Daten auch digitale Dateien, die komplexe Daten wie Text, Bilder, Video, Audio, .pdf-Dateien und branchenspezifische Dateiformate enthalten. Da Berge von unstrukturierten Daten durch Dinge wie schriftliche Dokumente erzeugt werden, ist das Versäumnis, diese wirksam zu verarbeiten, eine verpasste Chance, sich einen Wettbewerbsvorteil zu verschaffen. Es ist die Komplexität dieser unstrukturierten Datenquellen, die es extrem schwierig macht, sie zusammen mit anderen Datentypen zu analysieren. Ohne eine einzige Quelle, die alle Datentypen unterstützt, bleiben unstrukturierte Daten in Datensilos gefangen. Das führt dazu, dass Data Scientists unstrukturierte Daten nicht einfach durchsuchen, analysieren oder abfragen können, sondern sie aus verschiedenen Systemen zusammenführen müssen.

Neben den Problemen beim Datenmanagement ist es für Unternehmen faktisch unmöglich, alle zum Aufspüren von Geschäfts- und Wettbewerbstrends benötigten Daten selbst zu produzieren oder zu sammeln. Die Möglichkeit zur gemeinsamen Nutzung und Verknüpfung von Datasets – sowohl innerhalb eines Unternehmens als auch unternehmensübergreifend – wird zunehmend als der beste Weg angesehen, um mehr Wert aus Daten zu ziehen. Daher sind Data Scientists und Datenanalysten kontinuierlich auf der Jagd nach externen Daten, mit denen sie ihre ML-Modelle und Analysen untermauern und die Genauigkeit der Modellprognosen optimieren können.



Snowflake bietet Data Scientists und Datenanalysten Zugriff auf ein globales, vereinheitlichtes System zum Verwalten sämtlicher Arten von Daten, einschließlich unstrukturierter Daten. Angesichts der Zunahme an unstrukturierten Daten bietet die Snowflake Data Cloud auch in Zukunft eine einzige, konsolidierte Datenquelle, mit der Data Scientists und Datenanalysten den Zugriff auf Daten und deren Verarbeitung beschleunigen können, um aus allen Daten einen Mehrwert zu ziehen.

Hand in Hand mit dieser Fähigkeit, verschiedene Datentypen zu verwenden, ermöglicht Snowflake auch einen sicheren, kontrollierten und nahtlosen Zugriff auf externe Daten, sodass diejenigen, die Snowflake verwenden, ihre Compliance-Verpflichtungen im Rahmen verschiedener Datenschutzregelungen erfüllen können.

Die **Snowflake Data Cloud** ist ein Ökosystem, in dem Kunden und Partner von Snowflake sowie Datenanbieter und Datendienstleister eine Verbindung zu ihren eigenen Daten herstellen und Daten und Anwendungen von anderen Nutzern nahtlos gemeinsam nutzen und verwenden können. Mit der Snowflake-Plattform als Basis beseitigt die Data Cloud die durch Datensilos verursachten Probleme und bietet Unternehmen die Möglichkeit, eine einzige Kopie ihrer Daten zu bündeln und zu verbinden. Die Data Cloud stellt außerdem eine nahtlose Lösung dar, um über eine schnelle, einfache und kontrollierte Verbindung Nutzen aus den schnell wachsenden kommerziellen Datensets zu ziehen.

Die Data Cloud basiert auf der **Secure Data Sharing-Technologie von Snowflake**, die die traditionellen Barrieren bei der Datenübertragung beseitigt und Unternehmen so hilft, problemlos mit ihren internen und externen Geschäftspartnern zusammenzuarbeiten. Mit Snowflake werden Daten

niemals kopiert oder verschoben. Stattdessen gewährt der Anbieter den Benutzern Zugriff auf Live-Daten von ihrem ursprünglichen Standort aus. Dank der Trennung von Speicher und Rechenressource in Snowflake sind Latenzzeiten oder Konflikte durch gleichzeitige Benutzer kein Thema. Da Änderungen an den Daten an einer einzigen Version vorgenommen werden, sind die Daten für alle Verbraucher immer auf dem neuesten Stand und es wird sichergestellt, dass die Datenmodelle durchweg mit der neuesten Datenversion arbeiten.

Auf dem Snowflake Marketplace können Benutzer auf sofort abfragebereite Live-Daten und Anwendungen von Drittanbietern zugreifen und diese einkaufen. Durch die Nutzung derselben sicheren Technologie für Secure Data Sharing sind keine ETL-Prozesse erforderlich, sodass Data Scientists und Analysten Daten von Drittanbietern schneller nutzen können. Snowflake Marketplace vereinfacht und beschleunigt die Erkennung und Auswertung von Drittanbieter-Daten mit verwertbaren Datenbeispielen, die nahtlos mit Erstanbieter-Daten zur Validierung von Prototypen zusammengeführt werden können. Sobald ein Dataset ausgewertet wurde, kann der Einkauf direkt im Produkt abgewickelt werden und die zusätzlichen Kosten werden Teil Ihrer Snowflake-Rechnung.

Externe Daten stehen allen Data Cloud-Benutzern mit nur wenigen Klicks zur Verfügung. Sobald sich die Daten in der Data Cloud befinden, können sie freigegeben und genutzt werden. Es müssen weder CSV-Dateien gesendet noch manuelle Versionsprüfungen ausgeführt werden. Data Scientists können ihre Modelle um den nahtlosen Zugriff auf nahezu endlose Mengen an Daten zu jedem Thema erweitern – auch in Echtzeit und unter sich verändernden Bedingungen.

TREND 5: MIT PYTHON WERDEN AUCH DATA SCIENTISTS ZU APP-ENTWICKLERN

Data Scientists in jedem Unternehmen sind dafür verantwortlich, Daten und Modelle auf eine Weise freizugeben, die für ihre Kooperationspartner nachvollziehbar und überzeugend ist. Data Scientists sind jedoch oft gezwungen, ihre Ergebnisse in Dashboards darzustellen, die keine neuen Möglichkeiten zur Kommunikation dieser Informationen bieten. Außerdem haben Data Science-Teams selten Full-Stack-App-Entwickler zur Verfügung, um interne Produkte zu entwickeln, mit denen sie ihre Arbeit freigeben können.

Im kommenden Jahr wird sich der Status quo deutlich ändern, da Data Scientists dank Open-Source-Frameworks wie Streamlit auch als App-Entwickler in Python arbeiten können. Diese Frameworks erleichtern es, nur mit Python von der Idee zur App zu gelangen und so die Lücke zwischen ML und Aktion wirksam zu schließen.

Data Scientists können schnell interaktive Anwendungen mit reichhaltigen Komponenten wie Diagrammen, Eingabefeldern und vielem mehr entwickeln, um ihrem Team die Möglichkeit zu geben, mit Daten und Modellen zu arbeiten. Dies geschieht ohne die traditionelle Komplexität, die mit der Entwicklung einer Webanwendung verbunden ist, wie z. B. die Definition von Routen, die Bearbeitung von HTTP-Anfragen und das Schreiben von HTML, CSS oder JavaScript. Mithilfe dieser und ähnlicher Plattformen können Data Scientists Webanwendungen entwickeln, die ML-Modelle auf eine Weise zum Leben erwecken, die bisher für Data Science-Teams unerreichbar war.

Dank der Möglichkeit, nur mit Python in kürzester Zeit Anwendungen zu entwickeln und auf diesen interaktiven Plattformen freizugeben und weiterzuentwickeln, können Teams Daten leichter mobilisieren und ML-gestützte Erkenntnisse in die Hände von Geschäftsanwendern geben, damit diese Maßnahmen ergreifen können – und das ist ja schließlich das Ziel.

In der Vergangenheit war die Arbeit von Data Scientists und Geschäftsanwendern durch eine Kluft getrennt. Viele Data Scientists verbringen einen großen Teil ihrer Zeit und Energie mit der Entwicklung von Modellen, die für die Geschäftsanwender nicht unbedingt verständlich oder verwertbar sind. Ohne zugängliche, vertrauenswürdige Daten und Analysen gerät die Umsetzung ins Stocken, und das Versprechen von Data Science kann nie vollständig eingelöst werden.

Das alles ändert sich mit Streamlit. Streamlit wird bereits von vielen Fortune 500-Unternehmen eingesetzt und demokratisiert wirksam die Entwicklung von Webanwendungen. Und seit der Übernahme von Streamlit durch Snowflake gehören Problemfelder wie Infrastrukturmanagement, Elastizität und Security Data Governance für diese Anwendungen der Vergangenheit an. Im Rahmen einer fortlaufenden Integration können Benutzer ihre Anwendungen nahtlos bereitstellen und sicher gemeinsam nutzen und dabei auf die Sicherheit und Zuverlässigkeit der Infrastruktur von Snowflake bauen.

Je mehr Unternehmen Plattformen wie Streamlit einsetzen, um Data Scientists als App-Entwickler zu befähigen, desto schneller wird die Modellentwicklung, desto sinnvoller wird die Collaboration und desto effizienter werden Erkenntnisse ausgetauscht. Das begünstigt langfristiges Wachstum durch Agilität und Handlungsfähigkeit der Unternehmen.

TREND 6: KONTINUIERLICHES WACHSTUM BEI OPEN-SOURCE-ML-BIBLIOTHEKEN, -TOOLS UND -FRAMEWORKS

Data Science ist ein sich rasant entwickelnder Bereich. Es werden nicht nur jeden Monat neue Entwicklungen für ML und KI veröffentlicht – viele davon Open Source. Es erscheinen auch regelmäßig neue Start-ups, Tools und Lösungen auf dem Markt. Nehmen Sie zum Beispiel den rasanten Aufstieg und die atemberaubend schnelle Verbreitung von ChatGPT. Der von OpenAI entwickelte Chatbot mit einem fein abgestimmten Sprachmodell ging schnell viral und brachte zum ersten Mal Open-Source-KI in den öffentlichen Bereich. Angesichts der schnellen Entwicklung von Innovationen, der Akzeptanz und des Wandels in diesem Bereich ist es unumgänglich, sich nicht auf ein einziges Tool festzulegen.

Aus diesem Grund ist es ebenso wichtig, eine Plattform zu wählen, die von Frameworks und Algorithmen unabhängig ist und die auch mit anderen Tools zusammenarbeiten kann. Mit der Entscheidung für eine zukunftssichere Plattform stellen Sie sicher, dass auch alle künftigen ML-Tools nahtlos auf ihrer Plattform funktionieren. Schließlich möchten Sie keine neue Plattform einführen, nur um die Tools der nächsten Generation nutzen zu können.

Was die Plattform von Snowflake so einzigartig macht, ist ihre moderne Architektur. Snowflake wurde mit separaten, aber logisch integrierten Rechen- und Speichersystemen entwickelt. Dadurch entfällt der manuelle Aufwand für den Aufbau von Clustern, der bei anderen Systemen erforderlich ist, damit die einzelnen Ebenen zusammenarbeiten. So bietet Snowflake eine **Architektur mit mehreren Clustern und gemeinsam genutzten Daten**, die nahezu unbegrenzte Skalierbarkeit, unmittelbare Elastizität und ein extrem hohes Maß an gleichzeitigen Abläufen bietet, um die gesamte für Data Science und ML erforderliche Verarbeitung von der Datenaufbereitung bis zur Modellinferenz und der Bereitstellung in einer Anwendung zu ermöglichen.

Neben der zugrunde liegenden Architektur, die alle Datentypen unterstützt und mehrere Sprachen für die Verarbeitung in einer einzigen Engine zusammenführt, unterstützt Snowflake die Integration mit dem Data Science-Ökosystem auf vielfältige Weise.

- Mithilfe der **External Functions** von Snowflake können Benutzer mit jedem gehosteten oder benutzerdefinierten ML-Dienst eines Drittanbieters außerhalb von Snowflake über SQL interagieren. Dabei kann es sich z. B. um einen Voice-to-Text-KI-Dienst und andere NLP-Dienste handeln, oder sogar um ein ML-Modell, das in einer externen Umgebung für Echtzeit-Prognoseanfragen eingesetzt wird.
- Bei der Entwicklung und Bereitstellung von Python- und Java-Code mit **Snowpark** waren Entwickler schon immer flexibel und konnten mit ihrer bevorzugten integrierten Entwicklungsumgebung (IDE) oder ihrem Notebook arbeiten. Dank der Snowpark-Client-Bibliothek können Entwickler die Verarbeitung von nahezu jedem Entwicklungstool in Snowflake

übertragen, sodass Sie nicht auf die Verwendung der Tools verzichten müssen, die Sie bereits kennen.

- Da die Stärke von Python in einem umfangreichen Ökosystem von Open-Source-Paketen liegt, freuen wir uns, im Rahmen des Snowpark for Python-Angebots über unsere Anaconda-Integration nahtlose Open-Source-Innovationen auf Unternehmensebene in die Data Cloud zu bringen. Mit der umfassenden Sammlung von Open-Source-Paketen und der nahtlosen Verwaltung von Abhängigkeiten in Anaconda können Sie Ihre Python-basierten Arbeitsabläufe beschleunigen.

Snowflake verfügt außerdem über Integrationen mit beliebten Open-Source-Tools für die vielen Ebenen des Stacks für maschinelles Lernen. Dazu gehören Integrationen mit den beliebten Apache Iceberg-Tabellen (in Preview), um einen einfacheren Zugriff auf alle Ihre Daten zu erlauben, sowie Integrationen mit Feature Store-Lösungen wie Feast, um den gesamten Lebenszyklus der ML-Features zu verwalten.

- Die Zunahme unstrukturierter Daten geht einher mit der Entwicklung von Methoden zur Verwaltung und Verarbeitung dieser Daten. So können beispielsweise mit den neuen Kennzeichnungsdiensten Bilder und andere unstrukturierte Daten manuell getaggt werden. Mit **Snowflake Secure Data Sharing** können unstrukturierte Daten (in der öffentlichen Vorschau) für einen Anbieter freigegeben werden, der das Tagging der Daten übernimmt, ohne dass die Daten verschoben werden müssen. Zusätzlich können Analysetools auf die unstrukturierten Daten in Snowflake aufsetzen, um diese mit NLP-Diensten von Unternehmen wie Hugging Face und mit KI-Cloud-Diensten wie AWS Rekognition zu optimieren.

BESCHLEUNIGEN DES MASCHINELLEN LERNENS IN 2023

Es ist bemerkenswert, wie schnell Data Science zum Mainstream wurde. In den letzten 10 Jahren haben Unternehmen ihren Schwerpunkt vom Reporting und von historischen Analysen auf die Nutzung von Data Science mit fortgeschrittenen mathematischen Modellen und ML erweitert. Aber die meisten Unternehmen, die maschinelles Lernen einsetzen, müssen erst noch sehen, wie sich ihre Investitionen auszahlen, da sie damit kämpfen, ihre Projekte in die Produktionsphase zu bringen.

Eine moderne Datenplattform bildet eine notwendige Grundlage, um den Datenzugriff und die Datenverarbeitung in einer Weise zu gestatten, die leicht skalierbar ist und die strengsten Sicherheitsanforderungen selbst in den am stärksten regulierten Branchen wie Finanzdienstleistungen und Gesundheitswesen erfüllt. Snowflake bietet eine Architektur, die viele Programmiersprachen unterstützt, darunter SQL und Python, und die Datenkonsolidierung, effiziente Datenaufbereitung, einfachen Zugriff auf Open-Source-Bibliotheken und umfassende Integrationen in das Data Science-Ökosystem ermöglicht. Ihre Daten werden mobilisiert, sodass Sie umgehend von neuen Trends in den Bereichen Data Science und ML profitieren.

Mit Snowflake entfällt die Komplexität bei der Übernahme von Data Science und maschinellem Lernen in die Produktionsphase. Sind Sie bereit, Ihre maschinellen Lernprozesse zu beschleunigen und deren vollen Wert zu erschließen?





ÜBER SNOWFLAKE

Die Snowflake Data Cloud bietet jedem Unternehmen die Möglichkeit, seine Daten zu mobilisieren. Mithilfe der Data Cloud können Kunden Datensilos zusammenführen, Daten entdecken und sicher freigeben sowie verschiedene analytische Workloads ausführen. Wo auch immer sich Daten oder Benutzer befinden, Snowflake bietet eine einheitliche Datenlösung, die sich über mehrere Clouds und geografische Regionen erstreckt. Tausende von Kunden in zahlreichen Branchen nutzen die Snowflake Data Cloud und bringen so ihre Unternehmen voran. Darunter fallen auch 573 Unternehmen der Forbes Global 2000 (G2K) aus dem Jahr 2022 (Stand: 31. Januar 2023).

Erfahren Sie mehr unter [snowflake.com](https://www.snowflake.com)



© 2023 Snowflake Inc. Alle Rechte vorbehalten. Snowflake, das Logo von Snowflake und alle sonstigen hier erwähnten Namen von Produkten, Funktionen und Services von Snowflake sind eingetragene Marken oder Marken von Snowflake Inc. in den USA und anderen Ländern. Alle anderen erwähnten oder verwendeten Markennamen oder Logos dienen ausschließlich der Identifikation und können die Marken ihrer jeweiligen Eigentümer sein. Snowflake darf nicht mit diesen Eigentümern in Verbindung gebracht oder von diesen unterstützt oder gefördert werden.

QUELLENVERWEIS

¹ bit.ly/3lkt1qx