snowflake®

6 TENDANCES CLÉS EN MATIÈRE DE DATA SCIENCE ET D'ANALYSES DE DONNÉES POUR 2023

Comment le Data Cloud accélère le machine learning

CHAMPION GUIDES

EBOOK

TABLE DES MATIÈRES

3 Introduction

4 Focus sur des pratiques reproductibles de ML dans la production

5 Tendance n° 1 : outils et infrastructure unifiés pour SQL et Python

6 Tendance n° 2 : gestion et déploiement de fonctionnalités ML à grande échelle via des feature stores

7 Tendance n° 3 : les data scientists et analysts s'approprient la puissance des ML engineers

8 Tendance n° 4 : autres cas d'usage de données non structurées

10 Tendance n° 5 : les data scientists deviennent aussi des développeurs d'applications avec Python

11 Tendance n° 6 : croissance continue des bibliothèques, outils et environnements ML open source

12 Accélérez votre adoption du machine learning en 2023

13 À propos de Snowflake

INTRODUCTION

La data science s'est considérablement développée au cours des 10 dernières années. Ainsi, elle englobe désormais les domaines émergents de l'intelligence artificielle (IA) et du machine learning (ML) et a alimenté l'essor du Big Data, c'est-à-dire d'importants volumes de données arrivant à grande vitesse. Tandis que la data science étend sa portée et passe à l'échelle supérieure, le cloud computing suit le rythme et apporte des ressources quasi illimitées aux professionnels pour soutenir des projets complexes et ambitieux.

Malgré des investissements plus conséquents en data science et machine learning (ML), très peu d'entreprises tirent pleinement parti de l'impact commercial ou de l'avantage concurrentiel de leurs analyses avancées. Pourquoi ? Le passage à la production reste un obstacle important.

Les data engineers et scientists et les ML engineers travaillent souvent de manière isolée, avec des copies de données distinctes, ce qui conduit à une plus grande complexité et à un manque de solutions collaboratives. Les silos de données, où les données sont dispersées sur plusieurs systèmes et data lakes, ralentissent les data scientists, tandis que la complexité du traitement cantonne l'attention des entreprises sur la gestion de l'infrastructure plutôt que sur la création de valeur.

Toutefois, un changement se profile. Les récentes avancées technologiques sont en passe de révolutionner le mode de travail des data scientists et des data analysts. En 2023, six tendances sont susceptibles d'accélérer l'adoption du machine learning (ML) et de transformer l'approche d'analyse descriptive et de diagnostic des entreprises, expliquant ce qui s'est passé et pourquoi, en analyse prédictive et prescriptive, prévoyant plutôt ce qui va se produire et fournissant des pointeurs majeurs pour influencer l'avenir.

Dans cet eBook, vous découvrirez comment :

- Une infrastructure unifiée prenant en charge plusieurs langages de programmation permet aux data scientists, aux data engineers et aux data analysts de tirer pleinement parti du potentiel de chacun d'entre eux.
- Le Data Cloud de Snowflake permet de développer l'accès, le partage et l'utilisation de divers types de données, notamment les données non structurées (en preview), par le biais d'un écosystème sécurisé offrant un accès à des données tierces prêtes à l'emploi.
- Les feature stores permettent aux data scientists de gérer et de déployer des fonctionnalités de ML à grande échelle en offrant la reproductibilité, la découvrabilité et l'évolutivité requises.
- Les data analysts et les data scientists savent de mieux en mieux utiliser la puissance des systèmes autrefois réservés aux ML engineers, afin de faciliter et d'utiliser des processus de production plus avisés.
- Le développement dans Python d'applications web de plus en plus accessibles permet aux data scientists d'élaborer des modèles plus compréhensibles et exploitables.
- Le développement rapide des bibliothèques, outils et environnements open source démontre le besoin d'une solution pérennisant les investissements en matière de ML et de data science.

FOCUS SUR DES PRATIQUES REPRODUCTIBLES DE ML DANS LA PRODUCTION

Ces dernières années, d'énormes investissements ont été consacrés à la data science, l'IA et le ML, dans le but de dégager des rendements financiers plus conséquents, d'optimiser les processus et d'améliorer la résilience générale des entreprises. Selon une étude mondiale de McKinsey sur l'IA, l'adoption de l'IA a plus que doublé depuis 2017, et les entreprises qui investissent davantage dans l'IA prennent une longueur d'avance sur leurs concurrents. La valeur de la data science est une évidence ; en 2023, les entreprises s'intéresseront à l'optimisation et à la rationalisation de la production.

Les entreprises s'attacheront également à optimiser l'impact des initiatives en matière d'IA, de ML et de data science en affinant les processus, en modernisant les systèmes existants et en adoptant des ensembles d'outils cohérents pour tirer pleinement parti de ces technologies. Actuellement, on assiste à un effondrement rapide des barrières qui ont longtemps divisé la data science, les utilisateurs professionnels et les développeurs de différents langages de programmation, entre autres.

Les remarquables avancées réalisées en 2022 laissent deviner six tendances prometteuses pour la data science et le ML en 2023. Les outils et les technologies émergents peuvent accélérer le travail des data scientists, supprimer les silos de données et développer le traitement des données structurées comme non structurées. Les avancées actuelles dans ce domaine rendent le métier passionnant.

Le cloud est à la base de cette accélération. Les data scientists, les data engineers et les data analysts tirent parti des technologies cloud, qui offrent des ressources de traitement évolutives et quasi illimitées. En outre, le cloud permet d'éliminer les silos de données en consolidant les data lakes, les data warehouses et les datamarts pour fournir un partage et une analyse des données simples, rapides et sécurisés dans un emplacement unique afin de renforcer la collaboration entre les équipes.

En résumé, les données deviennent plus que jamais exploitables et de nouveaux outils se basent sur cette réalité pour apporter des changements significatifs. Ainsi, les entreprises sont en passe de mobiliser les données non seulement pour prédire l'avenir, mais également pour augmenter la probabilité de certains résultats par le biais de l'analyse prescriptive.

Voici six tendances qui définiront la data science en 2023 et accélèreront la progression de l'analyse des données vers le ML.

TENDANCE N° 1 : OUTILS ET INFRASTRUCTURE UNIFIÉS POUR SQL ET PYTHON

Avec l'augmentation de la quantité de données et du nombre d'applications, les data warehouses existants on-premise et autres ont entravé l'évolutivité. Pour y remédier, le secteur a décidé de copier les données dans des magasins d'objets dans le cloud et de mettre en place une infrastructure de traitement pour des langages de programmation tels que Python, SQL et Java. Cela s'est traduit par une gestion d'infrastructure complexe et une collaboration limitée entre les équipes.

SQL et Python sont les principaux langages du paysage moderne des données, et chacun d'entre eux présente des avantages uniques. La vitesse d'interrogation et d'agrégation des données de SQL est un atout considérable, tandis que la capacité de Python à gérer des analyses et des transformations complexes grâce à son riche écosystème open source est indispensable pour de nombreuses entreprises. Chacun des langages présente des avantages évidents, mais dans l'absolu, ils appartiennent à des mondes différents. Fonctionnant sur une infrastructure distincte, chaque langage est développé avec des outils différents, ce qui a longtemps empêché les data scientists de tirer parti du meilleur de chacun d'eux.

Ce manque d'interopérabilité a créé un profond effet de cloisonnement, empêchant les utilisateurs d'un langage de collaborer sur des analyses ou des flux de travail avec les utilisateurs de l'autre langage. Même pour ceux qui maîtrisent les deux langages, le passage d'un code à l'autre et l'hétérogénéité du processus d'interrogation de chaque langage peuvent être à la fois frustrants et chronophages.

Pourtant, le fossé entre SQL et Python se réduit grâce à des outils comme dbt, Hex et Snowpark de Snowflake, qui combinent les langages pour n'importe quelle tâche en tirant parti des avantages de chacun pour les différentes opérations d'analyse. Voici comment :

dbt : flux de transformation des données pour les équipes chargées des données suivant les meilleurs pratiques en matière d'ingénierie logicielle telles que la modularité, la portabilité, le processus CI/CD et la documentation dans leur entrepôt cloud. Alors qu'il s'agissait d'un flux de transformation SQL, en 2022, dbt a introduit un second langage, Python, pour répondre à la demande croissante de solutions transparentes permettant de basculer d'un langage à l'autre sur un même projet.

Hex : Hex est une plateforme moderne dédiée à l'analyse et à la data science qui permet de se connecter facilement aux données et de les analyser dans des notebooks collaboratifs SQL et Python. Elle permet également de partager le travail sous forme de cas d'usage et d'applications de données interactifs. Pour fournir une évolutivité de traitement quasi illimitée, son approche consiste à déplacer le processus de calcul vers l'entrepôt, plutôt que de charger l'ensemble des données dans un notebook.

Snowpark : Snowpark constitue un nouvel environnement de développement pour Snowflake. Il permet aux data engineers, data scientists et data developers d'écrire du code dans leur langage de prédilection et de l'exécuter dans Snowflake. La capacité de Snowflake à prendre en charge les interfaces de développement en SQL, Python, Java et autres, lui permet de basculer facilement d'un contexte à l'autre sans avoir à déplacer de données ni à configurer des clusters séparés.

L'utilisation d'outils capables d'unifier les langages de programmation est essentielle pour favoriser la croissance continue par le biais de la collaboration. En 2023, les data engineers, scientists et analysts n'ont plus à travailler de manière isolée faute d'un langage commun ; ils peuvent collaborer pour passer des données brutes aux informations. Au final, ce partage des connaissances permet de créer des projets plus souples et d'obtenir de meilleurs résultats à long terme.

TENDANCE N° 2 : GESTION ET DÉPLOIEMENT DE FONCTIONNALITÉS ML À GRANDE ÉCHELLE VIA DES FEATURE STORES

Lorsque les data scientists développent de nouveaux modèles de ML, ils doivent traiter les tâches fastidieuses de préparation des données et de création des fonctionnalités. Les fonctionnalités sont créées en alimentant et en préparant des colonnes de données dans un format spécifique pouvant être injecté dans les modèles de machine learning. Une fois les fonctionnalités générées pour un modèle, les data scientists doivent s'atteler à réécrire les mêmes fonctionnalités ou passer du temps à rechercher des fonctionnalités existantes à utiliser pour le prochain modèle.

Heureusement, 2022 a vu une forte augmentation de l'adoption des feature stores, un référentiel central qui permet d'améliorer la recherche, la collaboration et l'évolutivité des fonctionnalités de ML. Ils permettent aux data scientists de trouver rapidement des fonctionnalités transformées et prêtes à l'emploi, accélérant ainsi les phases d'expérimentation et de production. Les feature stores offrent de nombreux avantages. Ils optimisent notamment la collaboration entre les équipes par le biais de la réutilisation des travaux d'autres data scientists. Ils réduisent également le temps et les efforts requis pour déployer un modèle entraîné dans un environnement de production. En effet, les data scientists n'ont plus besoin de redéfinir ce qui est souvent un pipeline de données existant.

Aujourd'hui, les feature stores sont de plus en plus considérés comme la meilleure méthode d'amélioration des modèles de ML, car les équipes peuvent accéder plus facilement aux données optimisées et affinées pertinentes pour leurs modèles. Toutefois, créer un feature store n'est pas ce qu'il y a de plus facile. L'opérationnalisation des fonctionnalités pose un défi de taille, car la reproductibilité, la découvrabilité et l'évolutivité doivent être intégrées au feature store.

1. **La reproductibilité des modèles** nécessite la centralisation des fonctionnalités à un emplacement unique et la création de diverses versions des données et fonctionnalités. Les data scientists doivent être en mesure de remonter le temps pour découvrir les fonctionnalités et données utilisées pour entraîner un modèle. Les fonctionnalités doivent également être définies une fois, puis disponibles pour tous les cas d'usage futurs. Cela implique de mettre à jour les feature stores régulièrement et de gérer le contrôle des versions.
2. **La découvrabilité** nécessite d'extraire le processus de création de fonctionnalités de chaque instance de notebook individuelle et de le centraliser dans un référentiel unifié. Pour améliorer la collaboration et favoriser une réutilisation efficace du feature store, un catalogue doit être créé pour faciliter la découverte et la recherche de fonctionnalités.
3. **L'évolutivité** signifie que les éléments d'un feature store peuvent évoluer sans charge opérationnelle lourde au fur et à mesure que les cas d'usage de ML se développent. Un feature store doit pouvoir s'étendre de quelques centaines à des millions de fonctionnalités et continuer à servir efficacement les flux de travail d'entraînement et d'inférence. Plutôt que d'exécuter des pipelines en double pour l'entraînement et l'inférence, la centralisation du traitement permet d'accélérer l'accès aux fonctionnalités et de réduire le traitement redondant.

Snowflake fournit deux approches pour la création de feature stores, qui évitent de créer de nouveaux systèmes ou silos entre les data scientists et les data analysts.

La première approche consiste à s'appuyer sur une solution open source telle que Feast comme interface de feature store, tandis que Snowflake devient le magasin et le moteur des fonctionnalités. Les fonctionnalités sont conservées sur la même plateforme de données, avec l'aide d'outils d'ingestion, d'ELT et de catalogage existants. Avec cette solution, les entreprises maîtrisent la gestion des pipelines de données et l'interface sur laquelle les data scientists découvrent et parcourent les données et les fonctionnalités centralisées en un seul endroit évolutif.

La seconde approche consiste à utiliser une solution de feature store gérée rattachée à Snowflake. Les données client sont conservées dans leur forme brute et modélisée dans le Data Cloud de Snowflake, tandis que la transformation des données est exécutée dans Snowflake à l'aide de SQL ou Snowpark pour Python, et l'orchestration et la gestion des pipelines sont abstraites par le fournisseur de feature store. Les partenaires Snowflake dans ce domaine sont Tecton et Iguazio.

TENDANCE N° 3 : LES DATA SCIENTISTS ET ANALYSTS S'APPROPRIENT LA PUISSANCE DES ML ENGINEERS

Python gagne en popularité dans un grand nombre de secteurs ; il n'est pas seulement utilisé par les développeurs de ML, qui l'utilisent traditionnellement pour sa capacité à traiter une quantité massive de demandes de données, mais également par les data scientists pour créer des modèles et des systèmes fiables basés sur un code intuitif et flexible. La popularité de Python ne cesse de croître à mesure qu'il devient la *lingua franca* de la data science. Dans les faits, 70 % des développeurs de ML et des data scientists déclarent aujourd'hui utiliser Python.

Historiquement, il était difficile d'exécuter Python à l'échelle de l'entreprise en raison de la complexité de son infrastructure et des risques de sécurité liés à son écosystème open source. De ce fait, malgré son évolutivité et ses performances, ce langage était réservé aux développeurs de ML qui sont à l'aise avec la gestion d'infrastructures complexes et de correctifs de sécurité pour les applications stratégiques. Pourtant, les nouveaux environnements de développement simplifient la gestion de l'infrastructure pour Python (comme pour les autres langages) et facilitent ainsi plus que jamais l'utilisation de Python à grande échelle par les data scientists et analysts.

À mesure que Python se répand à travers les fonctions de données, les équipes de data analysts pourront tirer parti de sa rapidité d'exécution et de ses bibliothèques open source pour participer au nettoyage et à la structuration des données. Cette plus grande participation permettra de limiter les erreurs, d'améliorer la productivité et de développer de meilleures informations, pour une prise de décision de qualité.

Cependant, la collaboration croisée ne s'arrête pas aux équipes d'analyse de données : en effet, les data scientists qui sont familiers avec Python verront probablement leur rôle s'élargir. Au lieu de s'appuyer sur Python pour les travaux exploratoires, ils joueront un rôle de plus en plus actif dans les travaux liés à la production. Nous assistons ici à un rapprochement crucial entre le monde du développement et celui de la production grâce à l'accès à de puissantes infrastructures qui fonctionnent dans les deux, dotées d'une solide base de données offrant un accès ad hoc à des données situées n'importe où et d'une capacité de calcul pouvant être facilement augmentée à la demande.

La création de clusters n'entraînant aucun décalage, les data scientists seront en mesure de travailler à une plus grande échelle en ayant des informations sur la manière dont leurs modèles seront envoyés en production.

L'intégration d'outils AutoML dans des plateformes évolutives permettra également aux data scientists et analysts disposant d'une expertise limitée en machine learning de se former rapidement et de déployer des informations pertinentes alimentées par le ML. Les outils AutoML automatisent les tâches associées au développement et au déploiement des modèles de ML, traditionnellement exécutées uniquement par des experts en la matière, les data scientists. Ainsi, davantage d'utilisateurs de données sont en mesure d'automatiser tout ou partie du flux de travail ML, incluant la préparation des données, l'entraînement et la sélection de modèles, etc.

Ces outils font une énorme différence pour les data scientists et analysts en s'occupant des tâches fastidieuses (chargement, sélection, préparation et nettoyage des données) qui prenaient auparavant jusqu'à 80 % de leur temps, et qui n'en prennent plus que 45 %, selon une enquête menée par Anaconda auprès de data scientists et publiée par Datanami. Ainsi, la technologie AutoML augmente la productivité et laisse davantage de temps pour se consacrer à l'analyse.

À mesure que les outils AutoML deviennent plus évolutifs et transparents, leur taux d'adoption va probablement augmenter et permettre à un plus grand nombre d'acteurs de l'équipe chargée des données d'exploiter la puissance du ML dans la production.

TENDANCE N° 4 : AUTRES CAS D'USAGE DE DONNÉES NON STRUCTURÉES

Selon les prévisions de Statista, la quantité finale de données créées en 2022 était de 97 zettaoctets, un chiffre qui devrait augmenter rapidement dans les années à venir. D'ici 2025, les prévisions d'IDC, publiées par Analytics Insight, indiquent également que 80 % des données mondiales seront non structurées, ce qui devrait alerter les entreprises puisque seulement 0,5 % de ces ressources sont analysées aujourd'hui.[1]

Ce volume anticipé de données non structurées montre le besoin croissant pour les data scientists d'être en mesure d'analyser les données non structurées, parallèlement aux données structurées et semi-structurées (c'est-à-dire, les données avec certains éléments structurels non formatés de manière conventionnelle, tels que les journaux d'utilisateur et l'activité web fournis dans un format tel que JSON).

Malheureusement, les données non structurées comprennent des fichiers numériques contenant des données complexes, telles que du texte, des images, de la vidéo, de l'audio ainsi que des fichiers pdf et des formats de fichier spécifiques à certains secteurs. Des montagnes de données non structurées sont générées par des éléments tels que des documents écrits et l'incapacité à les traiter correctement est une opportunité manquée de bénéficier d'un avantage concurrentiel. C'est la complexité de ces sources de données non structurées qui rend extrêmement difficile leur analyse ainsi que d'autres types de données. En l'absence d'une source de données prenant en charge tous les types de données, les données non structurées restent bloquées dans des silos. Par conséquent, les data scientists ne peuvent pas rechercher, analyser ou interroger facilement les données non structurées et doivent alors les rassembler depuis plusieurs systèmes.

Outre les problèmes de gestion des données, il est quasiment impossible pour les entreprises de produire ou de collecter toutes les données requises pour dévoiler des tendances commerciales et concurrentielles. La capacité à partager et combiner des ensembles de données au sein des entreprises et entre les entreprises est de plus en plus considérée comme la méthode de valorisation optimale des données. C'est pourquoi les data scientists et les data analysts sont constamment à la recherche de données supplémentaires pour compléter leurs modèles de ML et leurs analyses avec des données externes, afin d'améliorer la précision des prédictions de modèles.

Avec Snowflake, les data scientists et les data analysts ont accès à un système mondial et unifié pour gérer tous les types de données, notamment les données non structurées. Face à l'augmentation des volumes de données non structurées, le Data Cloud de Snowflake continuera de fournir une source de données unique et consolidée, permettant aux data scientists et aux data analysts d'accélérer l'accès aux données et leur traitement afin d'extraire de la valeur de toutes les données.

Outre cette capacité à utiliser différents types de données, Snowflake offre également un accès sécurisé, gouverné et transparent à des données externes, permettant aux utilisateurs de Snowflake de remplir leurs obligations de conformité dans le cadre de différents régimes de protection des données.

Le Data Cloud de Snowflake est un écosystème au sein duquel les clients, partenaires, fournisseurs de données et fournisseurs de services de données de Snowflake se connectent à leurs propres données, et partagent et utilisent de manière transparente les données et les applications des autres utilisateurs. S'appuyant sur la plateforme Snowflake, le Data Cloud élimine les obstacles imposés par les silos de données, ce qui permet ainsi aux entreprises d'unifier leurs données et de se connecter à une copie unique de leurs données. En outre, le Data Cloud est une solution simple pour valoriser des ensembles de données commercialisés en croissance exponentielle avec un accès rapide, facile, et gouverné.

La technologie Secure Data Sharing de Snowflake, qui permet aux entreprises de collaborer facilement avec leurs partenaires commerciaux internes et externes en supprimant les barrières traditionnelles au transfert de données, renforce le Data Cloud. Avec Snowflake, les données ne sont jamais copiées ni déplacées.

Au lieu de cela, le fournisseur accorde aux utilisateurs le droit d'accéder aux données en direct depuis leur emplacement d'origine. Grâce à la séparation du stockage et du calcul de Snowflake, la latence ou le conflit dû à des utilisateurs simultanés n'est jamais un problème. Étant donné que les modifications sont apportées à une version unique des données, celles-ci sont toujours à jour pour tous les utilisateurs et garantissent que les modèles de données utilisent toujours la dernière version.

Sur la Marketplace Snowflake, les utilisateurs peuvent parcourir et acheter en direct des applications et des données prêtes à être interrogées auprès de fournisseurs tiers. Grâce à cette même technologie de partage sécurisé de données, les processus ETL ne sont pas nécessaires, ce qui permet aux data scientists et analysts d'exploiter plus rapidement des données provenant de fournisseur tiers. La Marketplace Snowflake simplifie et accélère la découverte et l'évaluation des données tierces avec des échantillons de données exploitables qui peuvent être associés de manière transparente aux données internes pour valider les prototypes. Une fois l'évaluation d'un ensemble de données terminée, le processus d'achat peut être géré directement dans le produit et les frais additionnels seront intégrés à votre facture Snowflake.

Les données externes sont disponibles et accessibles pour tous les utilisateurs du Data Cloud en quelques clics. Une fois dans le Data Cloud, les données sont prêtes à être partagées et utilisées. Il n'est pas nécessaire d'envoyer des fichiers CSV ou de gérer manuellement différentes versions. Les data scientists peuvent enrichir les modèles par le biais d'un accès transparent à des données quasi illimitées sur n'importe quel sujet, y compris dans des environnements en temps réel et évolutifs.

TENDANCE N° 5 : LES DATA SCIENTISTS DEVIENNENT AUSSI DES DÉVELOPPEURS D'APPLICATIONS AVEC PYTHON

Les data scientists de toute entreprise sont chargés de partager les données et les modèles d'une manière compréhensible et convaincante pour leurs collaborateurs. Pourtant, ils sont souvent contraints de présenter les résultats dans des tableaux de bord non compatibles avec les nouvelles méthodes de communication de ces informations. De même, il est rare que les équipes de data science disposent de développeurs d'applications généralistes pour créer des produits internes pour partager leur travail.

Au cours de l'année à venir, nous connaîtrons une remise en question importante du statu quo. En effet, les data scientists pourront désormais développer des applications en Python grâce à des environnements de développement open source tels que Streamlit, qui permettent de convertir facilement une idée en application uniquement avec Python, réduisant ainsi l'écart entre le ML et la réalisation.

Les data scientists peuvent rapidement créer des applications interactives avec des composants riches tels que des graphiques, des champs de saisie et bien plus encore, pour donner à leurs équipes les moyens d'interagir avec des données et des modèles, en les libérant de la complexité traditionnelle que représente la création d'une application web comme la définition de routes, la gestion des requêtes HTTP et le codage HTML, CSS ou JavaScript. Grâce à cette plateforme et à d'autres plateformes similaires, les data scientists peuvent développer des applications web qui mettent en action des modèles de ML avec une accessibilité sans précédent pour les équipes de data science.

La capacité à créer rapidement des applications uniquement avec Python, puis à les partager et les reproduire sur ces plateformes interactives, permet aux équipes de mobiliser leurs données et de mettre rapidement des informations de ML à disposition des utilisateurs professionnels pour leur prise de décision, ce qui constitue l'objectif final, après tout.

Historiquement, un gouffre sépare le travail des data scientists et celui des utilisateurs professionnels. Ainsi, de nombreux data scientists consacrent une grande partie de leur temps et de leur énergie à développer des modèles qui peuvent être, ou non, totalement intelligibles ou exploitables par les utilisateurs professionnels. L'absence de données et d'analyses accessibles et fiables enlise la situation et la promesse de la data science ne peut jamais être entièrement tenue.

Streamlit change la donne. Déjà adopté par de nombreuses entreprises du Fortune 500, Streamlit démocratise de fait le développement d'applications web. Et depuis que Streamlit a été racheté par Snowflake, les problèmes de gestion d'infrastructure, d'élasticité et de gouvernance des données de sécurité sur ces applications appartiennent au passé. Dans le cadre d'une intégration continue, les utilisateurs peuvent déployer en toute transparence et partager en toute sécurité leurs applications en s'appuyant sur la sécurité et la fiabilité de l'infrastructure de Snowflake.

Comme de plus en plus d'entreprises adoptent des plateformes telles que Streamlit pour permettre aux data scientists de développer des applications, le développement de modèles va s'accélérer, la collaboration va gagner en pertinence et les informations seront partagées de manière plus efficace. Les entreprises pourront ainsi faire preuve de souplesse et d'engagement, favorisant la croissance à long terme.

TENDANCE N° 6 : CROISSANCE CONTINUE DES BIBLIOTHÈQUES, OUTILS ET ENVIRONNEMENTS ML OPEN SOURCE

Le secteur de la data science évolue rapidement. Non seulement le ML et l'IA avancent chaque mois, un grand nombre de ces développements étant proposé en open source, mais de nouvelles start-up et solutions et de nouveaux outils émergent régulièrement. Prenons l'exemple de l'émergence et de l'adoption extrêmement rapides de ChatGPT. Développé par OpenAI avec un modèle de langage affiné, ce chatbot est rapidement devenu viral, amenant l'IA open-source pour la première fois dans le domaine public. Face au rythme rapide de l'innovation, de l'adoption et du changement que connaît ce secteur, il est impératif de ne pas se laisser enfermer dans l'utilisation d'un seul outil.

Pour cette raison, le choix d'une plateforme compatible avec différents environnements et algorithmes, pouvant également interagir de manière cohérente avec d'autres outils, est important. En choisissant une plateforme tournée vers l'avenir, vous avez la garantie que les outils de ML à venir continueront de fonctionner correctement sur la plateforme dont vous disposez. En effet, vous souhaiterez sans doute éviter d'avoir à changer de plateforme uniquement pour pouvoir utiliser la dernière génération d'outils.

L'architecture moderne de la plateforme Snowflake en fait une solution unique. Conçue avec des ressources de calcul et de stockage distinctes, mais logiquement intégrées, Snowflake élimine les efforts manuels de création de clusters nécessaires aux autres systèmes pour unifier les diverses couches de traitement. En conséquence, Snowflake a créé une **architecture de données partagées multi-cluster** offrant une évolutivité quasi illimitée, une élasticité instantanée et des niveaux de simultanéité extrêmement élevés pour alimenter tous les traitements nécessaires à la data science et au ML, de la préparation des données à l'inférence de modèle en passant par le déploiement dans une application.

Outre l'architecture sous-jacente qui prend en charge tous les types de données et rassemble plusieurs langages de traitement dans un seul moteur, Snowflake prend en charge les intégrations avec l'écosystème de data science de diverses manières.

- Les **External Functions** de Snowflake permettent aux utilisateurs d'interagir avec tout service de ML tiers, hébergé ou personnalisé en dehors de Snowflake avec SQL. Cela peut inclure un service d'IA de conversion de paroles en texte et d'autres services de NLP (traitement automatique du langage naturel), ou même un modèle de ML déployé dans un environnement externe pour des demandes de prédictions en temps réel.
- Pour développer et déployer leur code Python et Java avec **Snowpark**, les développeurs ont toujours eu la possibilité de travailler depuis leur environnement de développement intégré (IDE) de prédilection ou leur notebook. Grâce à la bibliothèque client de Snowpark, ils peuvent intégrer le traitement de pratiquement tous les outils de développement à Snowflake, et ainsi continuer à utiliser les outils qu'ils maîtrisent.
- Comme la puissance de Python réside dans son riche écosystème de packages open source, dans le cadre de l'offre Snowpark for Python, nous sommes heureux d'introduire une innovation open source transparente de niveau professionnel pour le Data Cloud via notre intégration Anaconda. Grâce à l'ensemble complet de packages open source d'Anaconda et à la capacité de gestion transparente des dépendances, vous allez accélérer vos flux de travail Python.

  Snowflake dispose également d'intégrations pour les outils open source les plus courants des nombreuses couches de la pile de machine learning. Cela inclut les intégrations dans les tables Apache Iceberg populaires (en preview) afin de simplifier l'accès à toutes vos données, et les intégrations aux solutions de feature store telles que Feast pour vous aider à gérer le cycle de vie de bout en bout de vos fonctionnalités de ML.
- La croissance des données non structurées entraîne le développement parallèle de méthodes de gestion et de traitement de ces données. Par exemple, de nouveaux services d'étiquetage permettent de baliser manuellement les images, ainsi que d'autres données non structurées. Avec **Secure Data Sharing de Snowflake**, les données non structurées (en public preview) peuvent être partagées avec un fournisseur afin d'ajouter des balises aux données sans avoir à déplacer ces dernières. En outre, il est possible d'utiliser des outils d'analyse des données non structurées en complément de Snowflake pour améliorer ces données à l'aide de services NLP fournis par des entreprises telles que Hugging Face et de services cloud d'IA tels qu'AWS Rekognition.

ACCÉLÉREZ VOS PROCESSUS DE MACHINE LEARNING EN 2023

La data science est devenue une composante incontournable avec une rapidité déconcertante. Au cours des 10 dernières années, les entreprises sont passées du reporting et de l'analyse historique aux processus de data science en utilisant des modèles mathématiques avancés et le ML. Cependant, la plupart des entreprises qui utilisent le machine learning n'ont pas encore eu de retour sur investissement, car elles peinent à faire passer leurs projets en production.

Une plateforme de données moderne est essentielle pour permettre l'accès aux données et leur traitement de manière évolutive et répondant aux besoins de sécurité les plus stricts des secteurs les plus réglementés, tels que les secteurs de la finance et de la santé. Snowflake offre une architecture compatible avec de nombreux langages de programmation, notamment SQL et Python, qui permet de consolider les données, de les préparer de manière efficace, ainsi que d'accéder facilement à des bibliothèques open source et des intégrations étendues dans l'écosystème de la data science. Vos données sont mobilisées, ce qui vous permet de tirer immédiatement parti des nouvelles tendances en matière de data science et de ML.

Avec Snowflake, finie la complexité de la mise en production de la data science et du machine learning. Êtes-vous prêt à accélérer votre processus de machine learning et à en tirer tous les bénéfices ?

À PROPOS DE SNOWFLAKE

Snowflake permet à chaque organisation de mobiliser ses données grâce au Data Cloud Snowflake. Les clients utilisent le Data Cloud pour réunir au même endroit leurs données silotées, analyser et partager en toute sécurité les données, et exécuter diverses charges de travail analytiques. Quel que soit l'endroit où se trouvent les données ou les utilisateurs, Snowflake offre une expérience unique qui s'étend sur plusieurs clouds et régions. Au 31 janvier 2023, des milliers de clients de nombreux secteurs, dont 573 des Forbes Global 2000 (G2K) de 2022, utilisent le Data Cloud Snowflake pour dynamiser leur activité. En savoir plus sur **snowflake.com**

CITATIONS

[1] bit.ly/3lkt1qx