LIVRE BLANC

QUALITÉ ET GOUVERNANCE DES DONNÉES

Six principes pour des données prêtes pour l'IA

Créer un socle de données fiables pour l'IA

Sommaire

Résumé	3
Introduction	3
Six principes pour des données prêtes pour l'IA	4
Trust Score pour l'IA	12
Socle de données de Qlik Talend pour l'IA	13
Conclusion	14

Résumé

- Ce document présente les six principes à appliquer pour vous assurer que vos données sont « IA compatibles ».
- Selon ces principes, elles doivent être variées, disponibles au bon moment, exactes, sécurisées, identifiables et faciles à utiliser par les moteurs d'IA.
- Vous découvrirez également ce qu'est le Trust Score pour l'IA, qui montre si vos données sont conformes à ces principes.

Introduction

Les secteurs de la santé, de la production et du service client peuvent s'attendre à grandement bénéficier de l'intelligence artificielle (IA), qui devrait leur donner les moyens d'offrir une meilleure expérience à leurs clients et employés. Les technologies d'IA comme le machine learning (ML) ont déjà permis aux utilisateurs de données de fournir des prévisions mathématiques, de générer des insights et d'améliorer la prise de décisions. Les technologies d'IA émergentes, comme l'IA générative, peuvent créer du contenu incroyablement réaliste, susceptible d'augmenter la productivité dans pratiquement toutes les activités de l'entreprise.

D'après Gartner, d'ici à 2025, l'IA générative sera largement utilisée dans 90 % des multinationales. De plus, à l'horizon 2026, plus de 80 % des entreprises s'appuieront sur des applications basées sur l'IA générative afin d'optimiser leurs processus de production[1].

Néanmoins, pour bénéficier des avantages de l'IA, vous devez vous assurer de la qualité de vos données. Découvrez, dans ce livre blanc, les six principes permettant de préparer vos données pour l'IA.

[1] « Analysts to Discuss Generative AI Trends and Technologies », Gartner, octobre 2023.

Six principes pour des données prêtes pour l'IA

Penser qu'il suffit d'injecter des données dans différents projets d'IA pour que la magie se produise serait un peu trop simpliste. Pourtant, c'est ce que font de nombreux utilisateurs de données. Au début, cette approche peut fonctionner, en apparence du moins, car les data scientists passeront vite de plus en plus de temps à corriger et à préparer les données à mesure que les projets se développeront.

Les applications intelligentes ont, en effet, besoin de données de qualité, correctement préparées, complètes, sans erreurs et organisées d'une manière facile à comprendre par les moteurs d'IA, ce qui suppose de longues heures de nettoyage et d'amélioration. Bien souvent, des informations supplémentaires sont également nécessaires, comme des définitions et des étiquettes, afin d'enrichir la signification sémantique pour l'apprentissage automatique et d'aider l'IA à effectuer des tâches plus efficacement.

Plus tôt les données destinées aux processus d'IA ultérieurs seront préparées, meilleurs seront les bénéfices que vous pourrez en tirer. Disposer de données prêtes pour l'IA revient à fournir à un chef cuisinier des légumes déjà lavés et émincés : le gain de temps et d'efforts permet d'obtenir plus vite un résultat satisfaisant. Le diagramme ci-dessous définit les six principes essentiels pour s'assurer que les données sont « prêtes » et adaptées à une utilisation avec l'IA.

Ils seront étudiés plus en détail dans les sections suivantes de ce livre blanc.

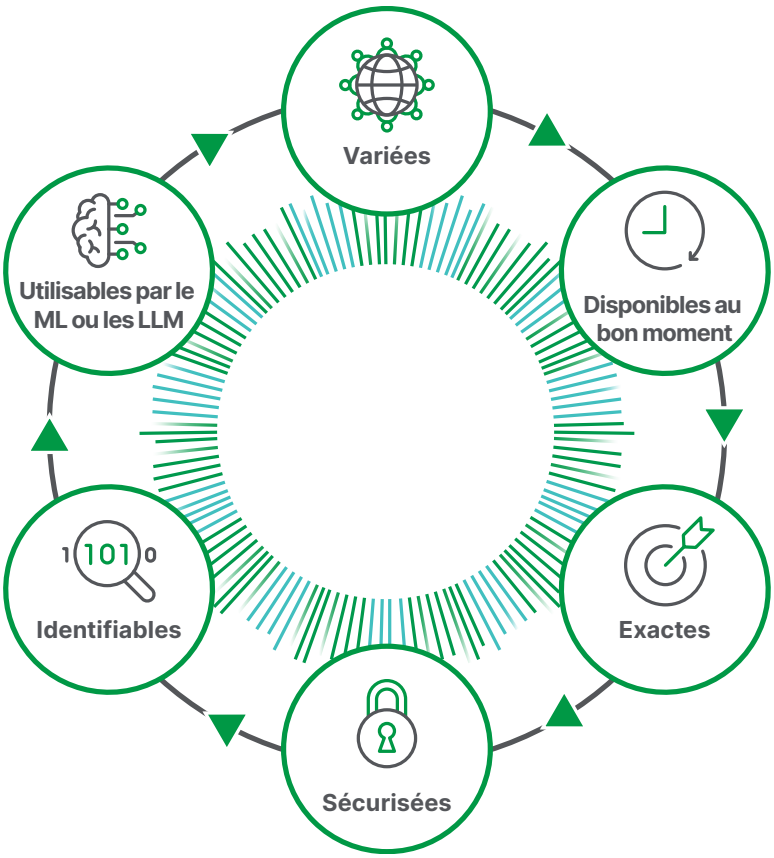

Schéma 1. Six principes pour des données prêtes pour l'IA

1 Les données doivent être variées

On considère que les systèmes d'IA sont biaisés (biais du machine learning ou de l'algorithme) lorsque les résultats des applications d'IA reflètent les biais humains comme les inégalités sociales. Ces biais apparaissent si des notions préjudiciables sont incluses lors du développement d'un algorithme ou, plus souvent, lorsque les données d'entraînement sont elles-mêmes biaisées. Un algorithme d'octroi de crédit qui fonctionne sur une plage de données financières restreinte peut ainsi refuser un prêt.

Notre premier principe consiste donc à fournir des données variées aux modèles d'IA. Cette diversité accrue limite les biais et les risques que les applications d'IA prennent des décisions inéquitables.

Avec des données variées, vos modèles d'IA ne sont pas conçus sur la base de datasets restreints et en silos, mais sur des sources diversifiées, couvrant des patterns, perspectives, variations et scénarios différents, bien que liés à un domaine spécifique. Ces données peuvent être bien structurées et hébergées dans le cloud ou on-prem. Elles peuvent aussi se trouver sur un mainframe, dans une base de données, un système SAP ou une application SaaS. Si elles ne sont pas structurées, elles peuvent être enregistrées dans des fichiers ou des documents conservés dans un espace de stockage d'entreprise.

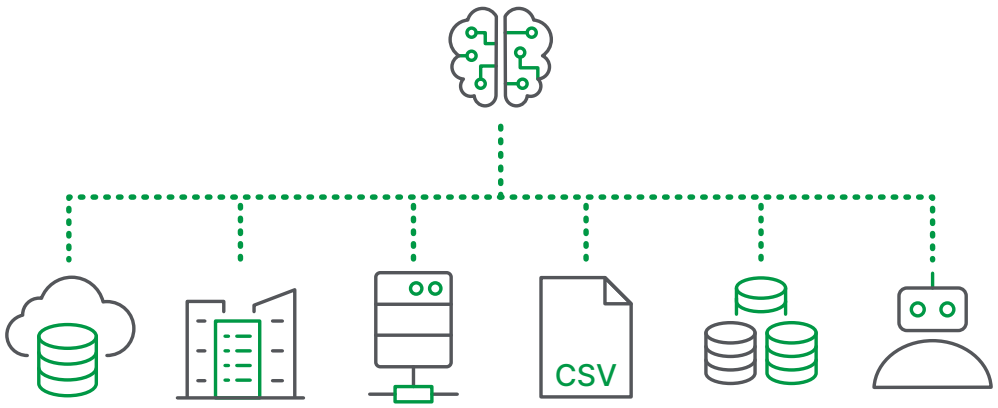

Schéma 2. Diversité des données

Les données à intégrer dans vos applications de ML et d'IA générative doivent absolument être variées et de différents formats.

2 Les données doivent être disponibles au bon moment

L'efficacité des applications de ML et d'IA générative dépend de la diversité des données, certes, mais aussi de leur récence. Si vous partez en excursion, le bulletin météo de la veille ne vous servira pas à grand-chose. Pour l'entraînement de modèles d'IA, c'est pareil : des informations obsolètes risquent de donner des résultats inexacts ou non pertinents. Avec des données récentes, les modèles d'IA peuvent suivre les tendances, s'adapter aux changements de situation et fournir des résultats optimaux. Ce qui nous mène au deuxième principe : la disponibilité des données.

Pour que vos données soient disponibles au bon moment, vous devez absolument créer et déployer des pipelines de données en temps réel et à faible latence. La **capture des changements des données (Change Data Capture, CDC)** est généralement choisie pour fournir au bon moment des données à partir de systèmes de gestion de bases de données relationnelles, tandis que la **capture en streaming** est utilisée pour des données provenant d'appareils IoT qui nécessitent un traitement à faible latence. Après la capture, les référentiels cibles sont actualisés et les **changements appliqués en continu** en temps quasi réel pour que les données soient les plus récentes possible.

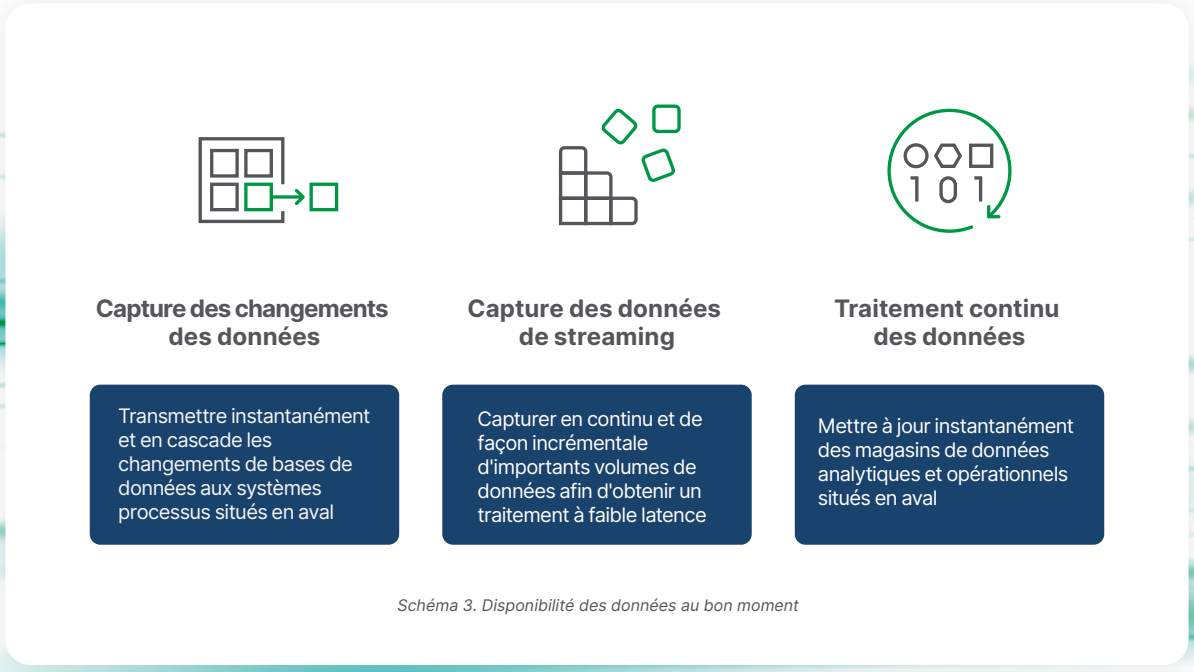

Schéma 3. Disponibilité des données au bon moment

N'oubliez pas que la disponibilité des données au bon moment améliore la précision et la pertinence des prévisions.

3 Les données doivent être exactes

L'exactitude des données est l'une des clés de la réussite d'un projet de ML ou d'IA générative, car les modèles d'IA sont comme des éponges sophistiquées qui apprennent et fonctionnent en se gorgeant d'informations. Si celles-ci sont fausses, c'est comme s'ils tentaient de nettoyer une surface avec de l'eau sale et le système d'IA produit des sorties biaisées et des créations sans queue ni tête, pour finir par dysfonctionner. L'exactitude des données, notre troisième principe, est donc un critère fondamental pour concevoir des applications d'IA fiables.

L'exactitude des données repose sur trois éléments. Le premier, le **profilage des données source**, permet de comprendre les caractéristiques, la complétude, la distribution, la redondance et la structure des données. On parle également d'analyse exploratoire des données (EDA).

Le deuxième est l'**opérationnalisation des stratégies de remédiation**, qui consiste à concevoir et déployer des règles de données, et à en surveiller continuellement l'efficacité. Vos data stewards peuvent intervenir à ce stade pour aider à dédupliquer et fusionner les données. L'IA peut aussi faciliter l'automatisation et l'accélération du processus grâce à des recommandations de qualité des données proposées par la machine.

Enfin, il faut permettre le lignage des données et l'analyse d'impact en fournissant aux data engineers et aux data scientists des outils qui montrent l'impact des changements potentiels de données et savent tracer l'origine de celles-ci pour éviter une modification accidentelle des données utilisées par les modèles d'IA.

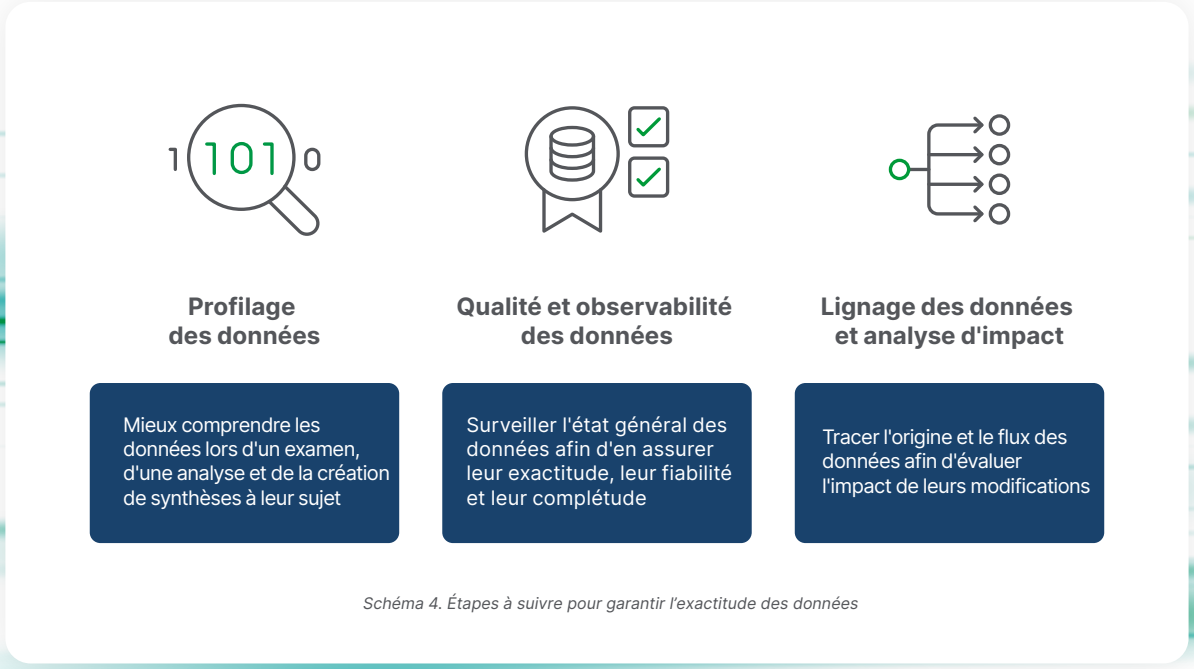

Schéma 4. Étapes à suivre pour garantir l'exactitude des données

Avec des données précises et de grande qualité, les modèles peuvent identifier les relations et les patterns pertinents, pour des décisions, une génération et des prévisions affinées.

4 Les données doivent être sécurisées

Les systèmes d'IA utilisent souvent des données sensibles, notamment des informations personnelles, financières ou propriétaires, qui doivent être manipulées de façon responsable. Intégrer des données dans une application d'IA sans les protéger revient à laisser ouverte la porte d'un coffre-fort. Des personnes malintentionnées pourraient voler les informations sensibles, manipuler les données d'entraînement pour produire des résultats biaisés ou même perturber le fonctionnement de systèmes d'IA générative entiers. Il est primordial de protéger les données pour garantir leur confidentialité, préserver l'intégrité des modèles et assurer le développement responsable d'applications d'IA performantes. Le quatrième principe pour des données prêtes pour l'IA est donc la sécurité.

Là encore, et comme il est virtuellement impossible de sécuriser manuellement les données à grande échelle, trois tactiques peuvent vous aider à le faire automatiquement. La **classification des données** détecte, catégorise et étiquette les données qui alimentent l'étape suivante. La **protection des données** définit les règles telles que le masquage, la tokenisation et le chiffrement pour l'obfuscation des données. Enfin, la **sécurité des données** définit des règles qui encadrent le contrôle d'accès (qui peut accéder aux données, par exemple). Ces trois concepts s'associent et fonctionnent ainsi : les niveaux de sécurité doivent d'abord être définis et les données étiquetées comme *sensibles*, *confidentielles* ou *avec restrictions d'accès*. Ensuite, une politique de protection doit être appliquée pour masquer les données avec restrictions d'accès. Enfin, il faut limiter les droits d'accès grâce à une règle de contrôle.

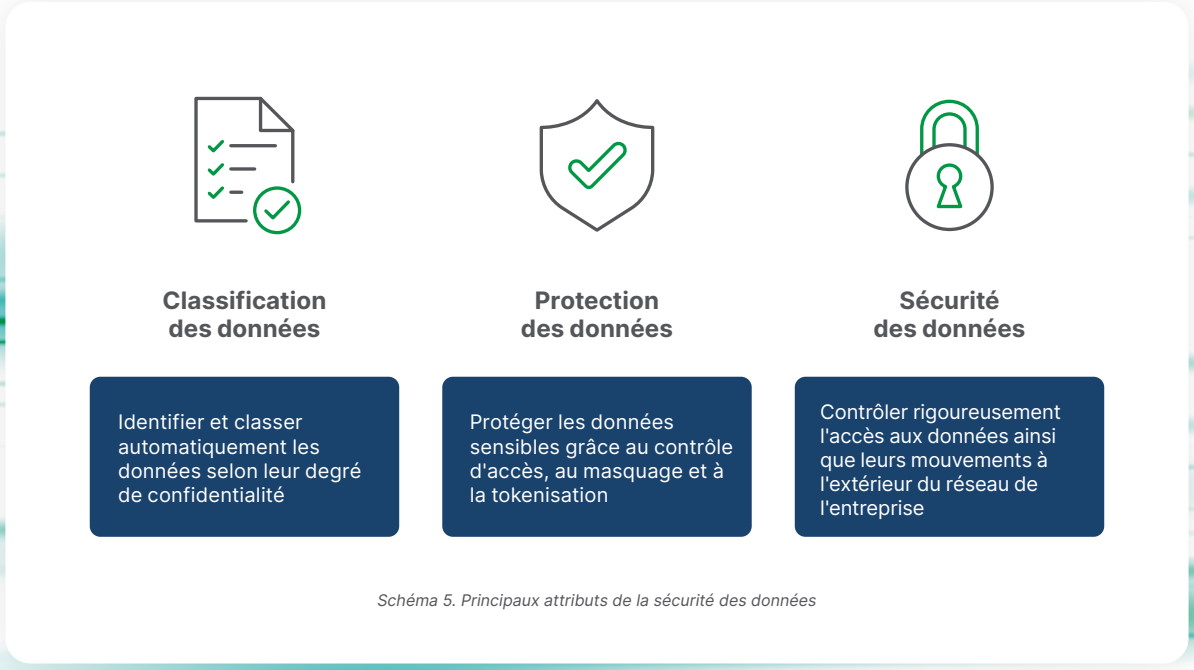

Schéma 5. Principaux attributs de la sécurité des données

Ces trois tactiques de protection sont essentielles pour renforcer la confiance globale dans votre système d'IA et préserver sa réputation.

5 Les données doivent être identifiables

Les principes présentés plus haut ont pour but premier de fournir rapidement et dans le bon format les données nécessaires aux personnes, applications d'IA ou systèmes cibles qui en ont besoin. Pour qu'elles soient prêtes pour l'IA, les données doivent être identifiables et accessibles dans le système. Les accumuler ne suffit pas. Imaginez une bibliothèque où tous les livres sont sous clé : le savoir est là, mais il est inatteignable. Les données identifiables permettent au ML et à l'IA générative de disposer des informations nécessaires pour apprendre, s'adapter et produire des résultats incroyables, libérant ainsi tout leur potentiel. Le cinquième principe est donc de disposer de données identifiables.

Sans surprise, les méthodes de création de métadonnées efficaces détermine la capacité à identifier les données. En parallèle des métadonnées techniques associées aux datasets IA, il faut aussi définir les métadonnées métier et le typage sémantique. Le **typage sémantique** produit un niveau de sens supplémentaire pour les systèmes automatisés, tandis que les termes métier donnent plus de contexte, ce qui facilite la compréhension humaine. Pour que les concepts soient compris par tous de la même façon, il est utile de créer un **glossaire** associant les termes métier aux éléments techniques dans les datasets. L'augmentation assistée par l'IA peut aussi être utilisée pour générer automatiquement la documentation et ajouter la sémantique métier issue du glossaire. Pour finir, un **catalogue de métadonnées** permet d'indexer toutes les métadonnées et d'y faire des recherches.

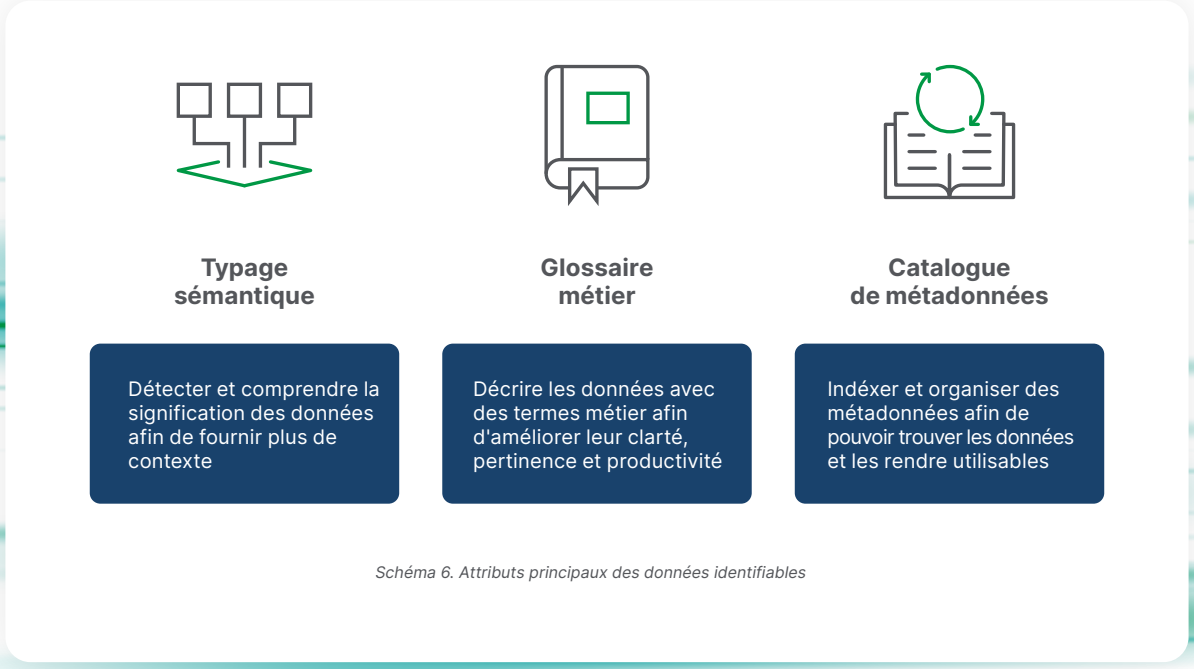

Schéma 6. Attributs principaux des données identifiables

Cette approche garantit que les données sont directement identifiables, appliquables, fonctionnelles et pertinentes pour la tâche d'IA concernée.

6 Les données doivent être facilement utilisables par les systèmes de ML ou de LLM

Comme nous l'avons vu, les applications de ML et d'IA générative sont des outils dont le potentiel dépend de la capacité à consommer facilement les données. Les humains savent déchiffrer une note manuscrite ou se repérer dans une feuille de calcul brouillonne, mais pour les technologies d'IA, les informations doivent être présentées dans un format spécifique. Imaginez : si une personne difficile ne mange pas ce que vous lui servez, elle aura faim. Pour les initiatives d'IA, c'est la même chose : elles échoueront si le format des données destinées à une application de ML ou de large language model (LLM) n'est pas le bon. Faciliter la consommation des données libère le potentiel de ces systèmes d'IA, et leur permet d'ingérer sans problème les informations et de les traduire en actions intelligentes pour obtenir des résultats créatifs. Le sixième et dernier principe pour des données prêtes pour l'IA est donc de faciliter la consommation des données.

Faciliter la consommation des données pour le machine learning

Sans transformation des données, pas de consommation des données pour le ML. Si les algorithmes tels que la régression linéaire tiennent le devant de la scène, la qualité et la forme des données qui les entraînent sont tout aussi importantes. De plus, les efforts nécessaires pour nettoyer et organiser les données, et ainsi favoriser leur consommation par les modèles de ML, sont largement récompensés. Les données préparées permettent d'entraîner efficacement les modèles et d'obtenir des prévisions précises et des résultats fiables, pour une réussite globale et concrète du projet de ML.

Toutefois, le format des données d'entraînement dépend largement de l'infrastructure de ML sous-jacente. Les systèmes de ML traditionnels sont basés sur des disques, et les data scientists travaillent surtout à mettre en place des bonnes pratiques et des procédures de codage manuel pour gérer de grands volumes de fichiers. Récemment, des systèmes de ML basés sur un lakehouse ont commencé à utiliser un magasin de fonctionnalités de type base de données et les data scientists sont passés au langage SQL, le plus utilisé. Les structures en tableaux de données bien formées et de grande qualité sont le format le plus pratique et le plus facile pour les systèmes de ML.

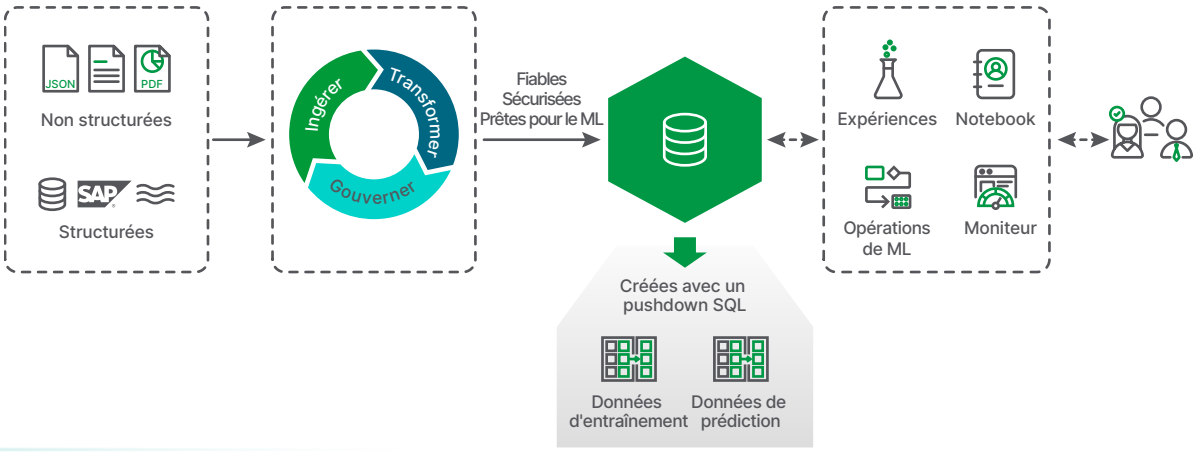

Schéma 7. Faciliter la consommation des données pour le ML

Faciliter la consommation des données pour l'IA générative

Les LLM comme GPT-4 (OpenAI), Claude (Anthropic), LaMDA et Gemini (Google AI) ont été pré-entraînés sur d'énormes volumes de données texte et sont au cœur de l'IA générative. On estime que le modèle GPT-3 d'OpenAI a été entraîné avec environ 45 To de données, soit plus de 300 milliards de tokens. Malgré ce nombre impressionnant d'entrées, les LLM ne savent pas répondre à des questions spécifiques sur votre activité, car ils n'ont pas accès aux données de votre entreprise. La solution consiste à les développer avec vos propres informations, ce qui permet d'obtenir des applications d'IA plus pertinentes et plus fiables.

On nomme « génération augmentée de récupération (RAG) » le mode d'intégration de vos données métier dans une application de LLM. Cette technique utilise généralement des informations texte dérivées de sources sur fichiers non structurées : présentations, archives de messagerie, documents texte, PDF, transcriptions, etc. Le texte est ensuite scindé en « chunks » faciles à gérer, puis converti en une représentation numérique utilisée par le LLM dans un processus nommé « embedding ». Ces embeddings sont ensuite stockés dans une base de données vectorielle comme Chroma, Pinecone ou Weaviate. Il est intéressant de noter que de nombreux fournisseurs de bases de données traditionnelles (PostgreSQL, Redis et SingleStoreDB, par exemple) prennent également en charge les vecteurs. De plus, les plateformes cloud comme Databricks, Snowflake et Google BigQuery viennent également de s'y mettre.

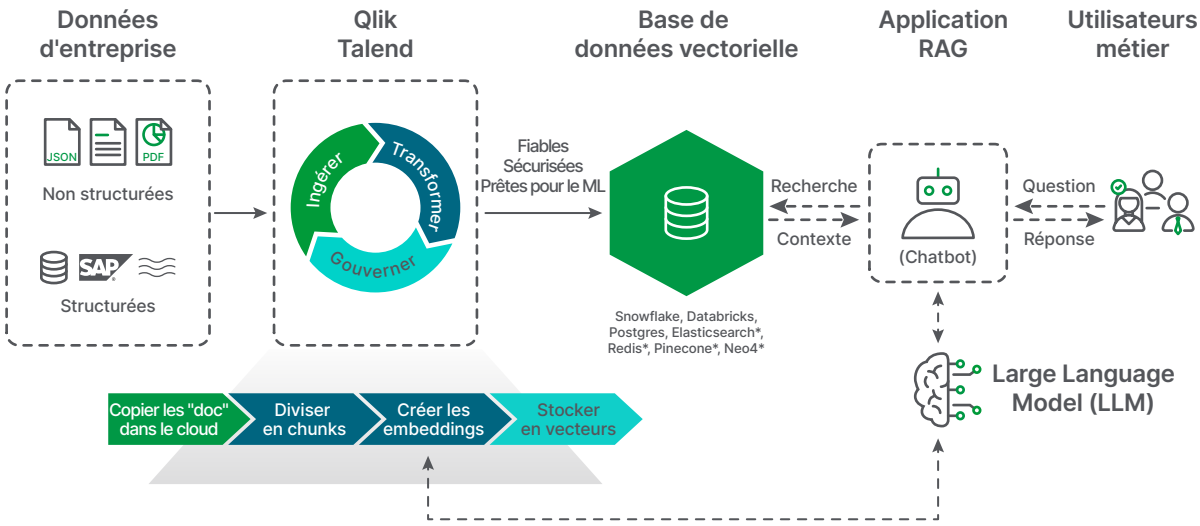

Schéma 8. Faciliter la consommation des données pour l'IA générative

Que vos données source soient structurées ou non, l'approche de Qlik permet de s'assurer que des données de qualité peuvent être facilement consommées par vos applications d'IA générative, de RAG ou de LLM.

Trust Score pour l'IA

Maintenant que les six principes fondamentaux pour préparer et rendre les données utilisables sont posés, reste à savoir s'ils peuvent être codifiés et facilement traduits pour un usage quotidien, et comment déterminer rapidement si les données sont prêtes pour l'IA. Pour cela, on peut utiliser le Trust Score pour l'IA de Qlik comme un indicateur compréhensible de préparation globale.

Le Trust Score pour l'IA attribue une dimension séparée à chaque principe, puis en agrège les valeurs pour générer un score composite, qui est un raccourci rapide et fiable évaluant le niveau de préparation de vos données pour l'IA. Comme les données d'entreprise changent sans arrêt, ce score est régulièrement vérifié et souvent recalculé pour suivre les tendances de préparation des données.

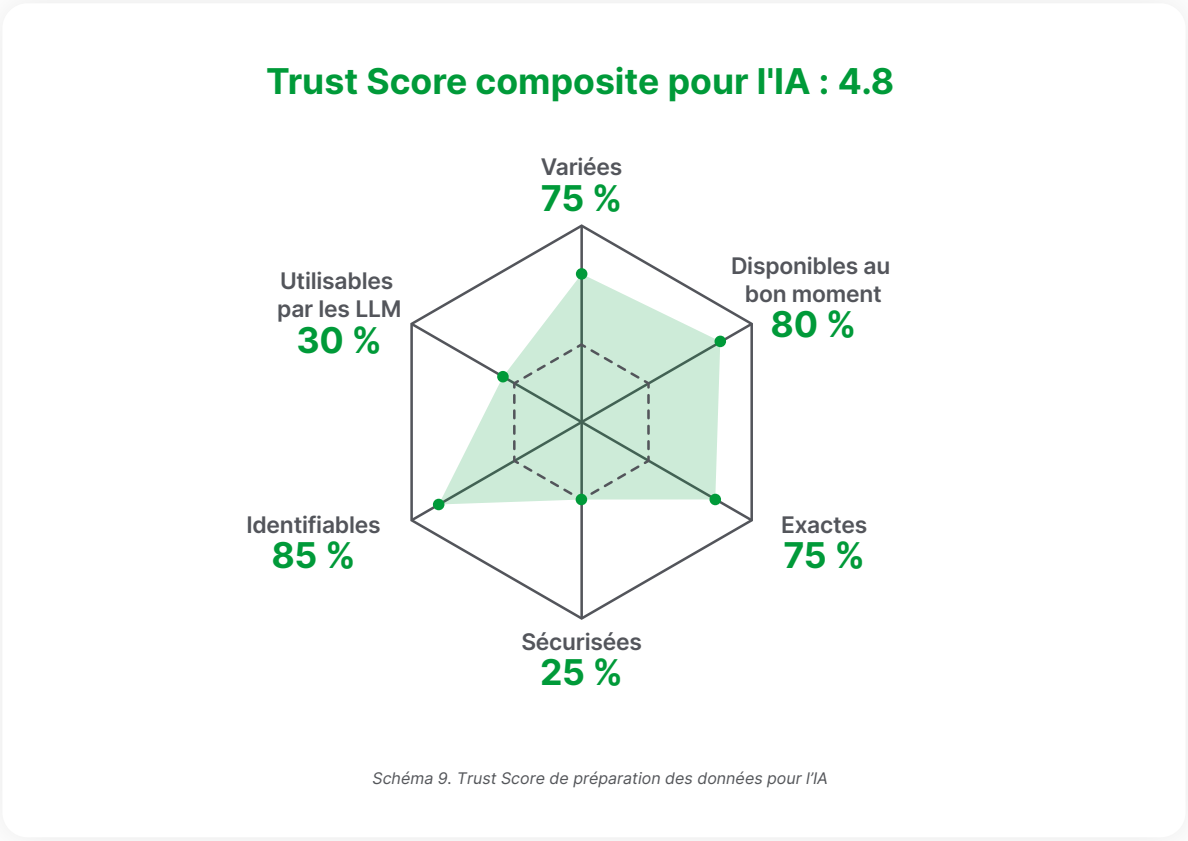

Schéma 9. Trust Score de préparation des données pour l'IA

Le Trust Score pour l'IA agrège plusieurs metrics dans un indice de préparation unique et facile à comprendre.

Socle de données de Qlik Talend pour l'IA

Jamais le besoin de données en temps réel et d'excellente qualité n'a été aussi important pour prendre des décisions plus réfléchies, améliorer l'efficacité opérationnelle et favoriser l'innovation métier. Des organisations performantes choisissent les solutions leaders d'intégration et de qualité des données de Qlik Talend pour livrer efficacement des données fiables aux warehouses, aux lakes et à d'autres plateformes de données d'entreprise. Nos solutions sophistiquées complètes utilisent des pipelines automatisés, des transformations intelligentes et des datasets fiables et de qualité, qui offrent aux professionnels de la data l'agilité nécessaire pour répondre aux exigences des organisations en matière de gouvernance et de conformité.

Qlik Talend peut donc vous accompagner dans la création de warehouses ou de lakes pour obtenir des analyses pertinentes, la modernisation de vos infrastructures de données opérationnelles pour améliorer votre efficacité ou l'utilisation de données multicloud pour les initiatives d'IA.

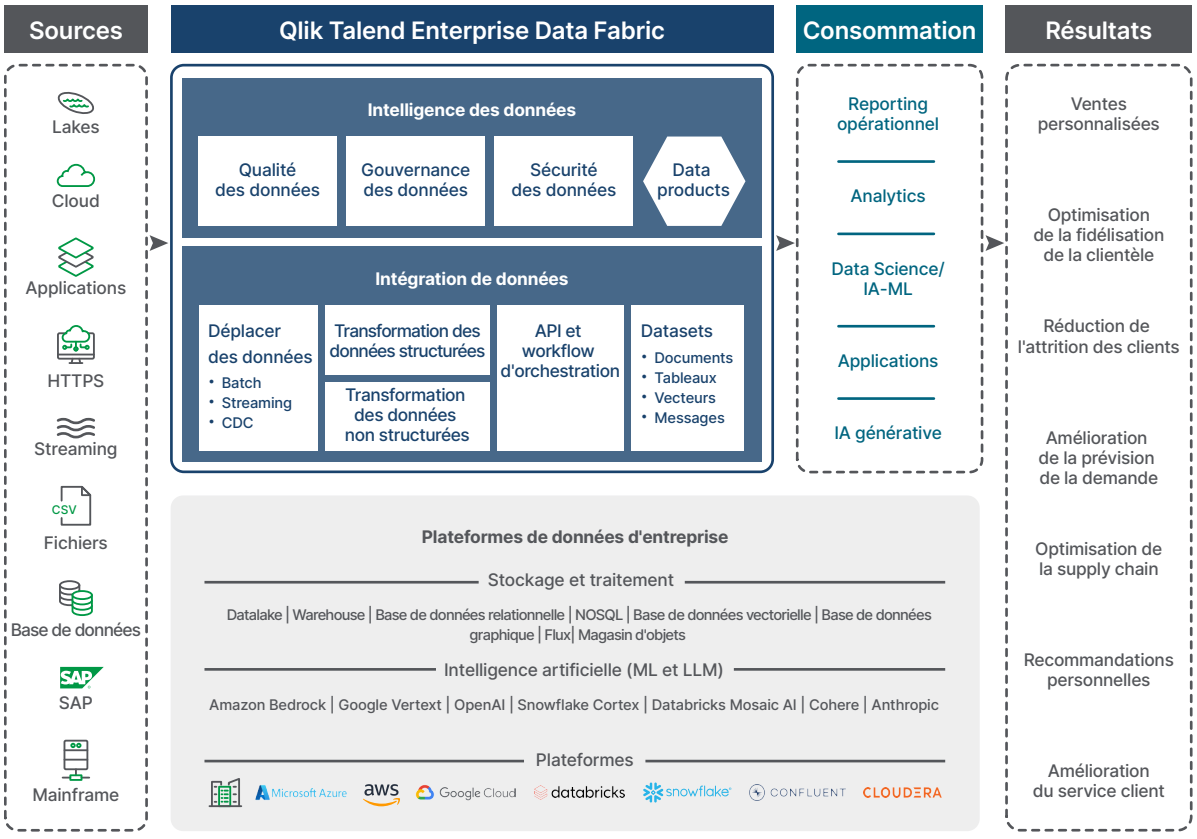

Schéma 10. Qlik Talend Enterprise Data Fabric pour l'IA et l'analytics

Conclusion

Au-delà de la puissance de transformation du machine learning et du potentiel de croissance extraordinaire de l'IA, la préparation des données reste fondamentale pour réussir l'implémentation de l'IA. Dans ce livre blanc, nous avons donc décrit six principes essentiels pour créer un socle de données solide et fiable. L'application combinée de ces principes aidera votre entreprise à exploiter tout le potentiel de l'intelligence artificielle.

Envie de révéler le véritable potentiel IA de votre entreprise ?

Se lancer

À propos de Qlik®

Qlik transforme les données issues d'environnements complexes en insights exploitables pour permettre aux entreprises d'atteindre leurs objectifs stratégiques. Plus de 40 000 clients dans le monde utilisent notre portefeuille de produits afin de profiter d'une qualité de données pervasive et de capacités d'IA/ML avancées, conçues pour les entreprises. L'intégration et la gouvernance des données sont nos domaines d'excellence. Cette expertise nous permet de proposer des solutions complètes qui supportent un large panel de sources, aussi hétérogènes soient-elles. Les analyses intuitives, en temps réel de Qlik mettent à jour des modèles cachés et donnent ainsi la possibilité aux équipes de relever des défis complexes et saisir de nouvelles opportunités. À la fois pratiques et évolutifs, nos outils d'IA/ML favorisent une prise de décisions plus éclairées, et ce plus rapidement. Notre technologie, compatible avec toutes les plateformes, et notre expertise sont un atout compétitif pour nos clients, faisant de Qlik un véritable partenaire stratégique.

qlik.com

24-DQ-0001-01-WHM