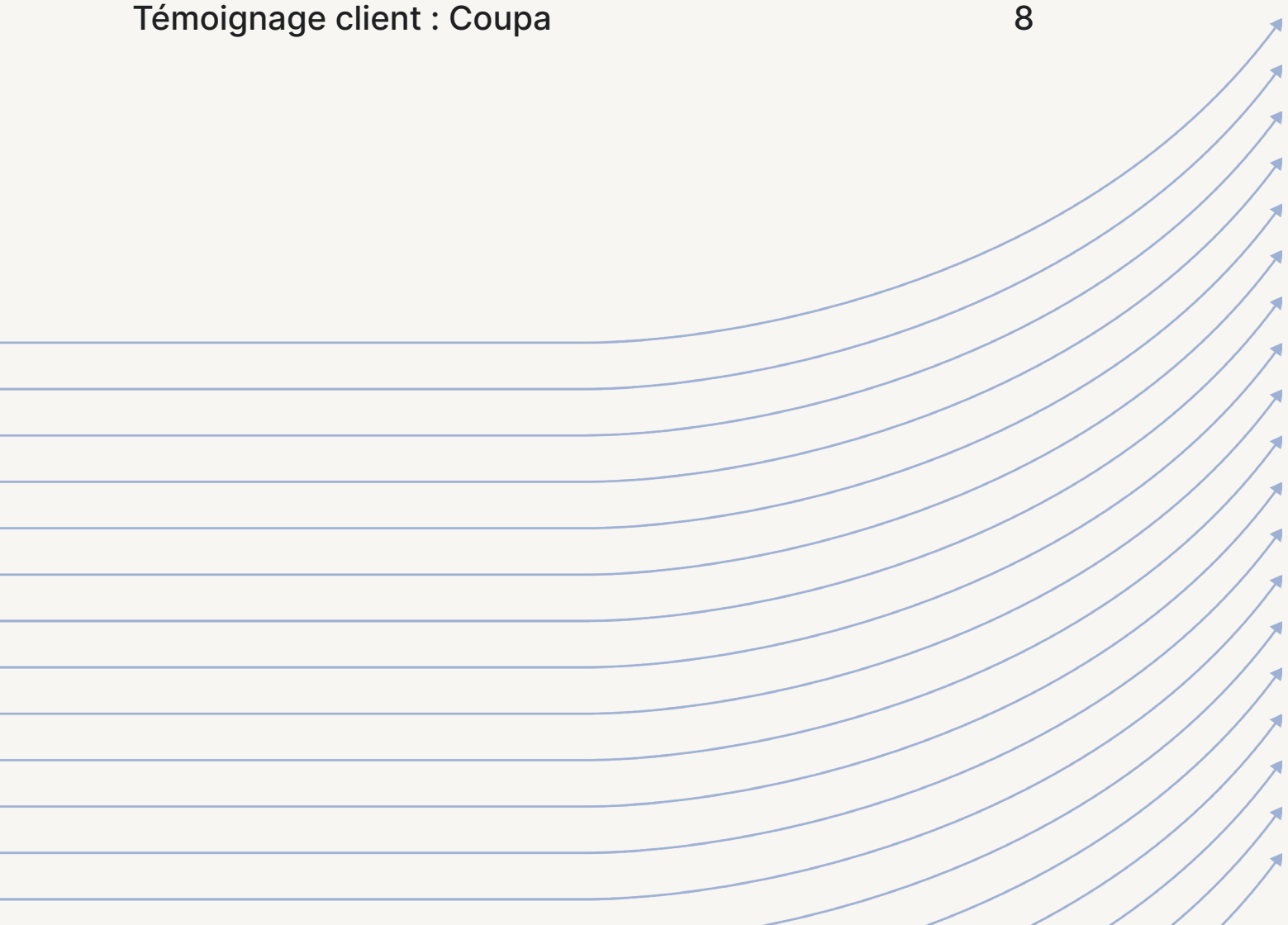
Fivetran

Guide du CIO **sur la gestion des data lakes pour l'IA générative**

Comment créer un socle de data prêt pour l'analyse afin d'assurer la réussite de l'IA et de l'IA générative.

Table des matières



Synthèse

L'IA générative est prête à révolutionner la productivité et la croissance des entreprises. McKinsey prévoit qu'elle pourrait ajouter des milliers de milliards à l'économie mondiale. Pourtant, une nouvelle étude du MIT révèle que si 82 % des cadres dirigeants et des cadres supérieurs accordent la priorité au déploiement des cas d'utilisation de l'IA, nombreux sont ceux qui rencontrent des difficultés pour préparer les data pour l'IA et l'IA générative.

La principale difficulté, citée par 45 % des cadres dans l'étude du MIT, concerne l'intégration des data et les pipelines. Ces chiffres sont préoccupants étant donné que les organisations consacrent en moyenne 13 % de leur chiffre d'affaires annuel aux initiatives en matière d'IA. Le fait que les dirigeants rencontrent des difficultés avec un aspect aussi fondamental de l'ingénierie data suggère qu'ils ne sont probablement pas prêts pour le déploiement et l'évolutivité de l'IA.

De nombreuses organisations sont confrontées à des pipelines de data créés manuellement qui consomment beaucoup de temps et de ressources, ce qui conduit à une inefficacité et entrave l'évolutivité de l'IA. En outre, sans gouvernance, les data lakes deviennent des « marécages de data » troubles et inutilisables au lieu d'être des data lakes fonctionnels. Le coût d'opportunité lié au report de la modernisation est plus élevé que jamais, car l'IA générative offre des gains d'efficacité importants.

Ce Guide électronique fournit aux CIO (directeurs de l'information) des stratégies pour optimiser et garantir la réussite des initiatives d'IA et d'IA générative grâce à un service de gestion des data lakes. Il met l'accent sur l'importance stratégique des data lakes, le rôle critique de la gouvernance data et la manière de rationaliser les opérations grâce à des solutions modernes de gestion des data.

> « Là où [Honeywell] a récemment déployé l'IA générative, l'entreprise constate déjà une augmentation considérable de la productivité ».
>
> **- Suresh Venkatarayalu, Chief Technology and Innovation Officer**
>
> **Source : "AI readiness for C-suite leaders" by MIT Technology Review Insights**

Le rôle du CIO dans la gestion des data lakes

En tant que dirigeants stratégiques, les CIO doivent aligner les stratégies data sur les objectifs de l'entreprise. Les CIO doivent :

Définir des missions qui relient les stratégies data aux cas d'utilisation utiles de l'IA/ML. Expliquer comment un environnement de data gouvernées et prêtes pour l'analyse catalyse l'innovation grâce à des initiatives d'IA et d'apprentissage automatique.

Gérer efficacement les coûts en normalisant les plateformes et en maîtrisant la prolifération des infrastructures. Éviter la prolifération d'infrastructures et d'outils multiples. La normalisation des plateformes simplifie les opérations et garantit un déploiement plus fiable de l'IA.

Faciliter un changement de culture en faveur de la gestion des data et d'une gouvernance solide dans tous les services. Si la technologie est essentielle, les CIO doivent également favoriser le changement organisationnel nécessaire à la transformation de l'IA.

La création d'un data lake bien géré et gouverné est essentielle tant sur le plan technologique que stratégique. Selon le MIT, 83 % des cadres dirigeants ont souligné la nécessité de consolider les nombreuses sources de data cloisonnées, une tâche qui exige le leadership du CIO pour une intégration et un mouvement efficaces des data.

Les data lakes évoluent vers des ressources multi-locataires qui prennent en charge des cas d'utilisation plus larges. Ils ont besoin de data fiables et conformes qui sont automatiquement cataloguées, nettoyées et préparées. Cette approche rationalisée permet non seulement de réduire le temps d'obtention des informations, mais aussi de stimuler l'utilisation active du data lake. En se concentrant dès le départ sur le déplacement, la qualité et la gouvernance des data, les CIO peuvent optimiser le développement de l'IA et réduire les risques.

La valeur stratégique des data lakes pour l'IA et l'apprentissage automatique (ML)

Les data lakes sont essentiels pour exploiter le potentiel de l'IA générative, car ils offrent des avantages significatifs par rapport aux bases de données et aux data warehouses traditionnels. Les data lakes :

Prennent en charge divers types de data nécessaires à l'entraînement de modèles sophistiqués d'IA et de ML. Contrairement aux systèmes traditionnels, les data lakes stockent et gèrent de grands volumes de data non structurées, semi-structurées et structurées sans traitement préalable important.

Préservent les data dans leur format brut, ce qui offre une flexibilité maximale pour de futurs cas d'utilisation analytique. En préservant les détails originaux et l'intégrité des data, les entreprises bénéficient d'une flexibilité maximale pour s'adapter à tous les cas d'utilisation analytique futurs.

Permettent l'évolutivité horizontale et verticale, en hébergeant de grandes quantités de data et des capacités de traitement robustes. Le déploiement horizontal permet aux entreprises de stocker des quantités pratiquement illimitées (pétaoctets) de data, tandis que le déploiement vertical permet un traitement et une analyse de haute performance sur de grands ensembles de data.

Offrent une ingestion flexible des data provenant de diverses sources, ce qui permet de créer des modèles d'IA précis et de prendre des décisions éclairées. La capacité d'intégrer des data structurées, semi-structurées et non structurées provenant de diverses sources permet aux data lakes de tout prendre en charge, des expériences à petite échelle aux déploiements d'IA d'entreprise.

Favorisent une analyse unifiée. En consolidant les data provenant de sources disparates, les data lakes fournissent la vision globale indispensable à la création de modèles d'IA précis. L'environnement de data unifié permet aux modèles d'IA de générer des informations transversales et de prendre des décisions éclairées. Un socle de data consolidé alimente en fin de compte des flux de travail d'analyse avancée et d'apprentissage automatique qui permettent d'obtenir de meilleurs résultats et des avantages stratégiques.

Les data lakes apportent de l'agilité tout en simplifiant la consolidation des data et en améliorant la précision des modèles d'IA. Pour tirer parti de l'IA générative, les CIO doivent prendre l'initiative d'architecturer des data lakes optimisés pour la réussite de l'IA. Les entreprises ont besoin d'une infrastructure de data à l'épreuve du temps, capable de s'adapter à l'évolution de leurs besoins.

La gouvernance data pour les data lakes

Soixante pour cent des cadres dirigeants affirment que la résolution des problèmes de gouvernance, de fiabilité et de sécurité des data est une condition préalable à la réalisation de leurs objectifs en matière d'IA. Une gouvernance data efficace est essentielle pour optimiser les initiatives d'IA qui se fondent sur les data lakes. Des cadres de gouvernance solides garantissent :

- **La qualité des data :** Des data propres, exactes et fiables produisent des modèles d'IA fiables.
- **La conformité :** Le respect des réglementations strictes en matière de confidentialité et de sécurité des data, telles que le GDPR et la CCPA, atténue les risques d'amendes et d'atteinte à la réputation.
- **La sécurité :** La mise en œuvre de mesures robustes telles que le cryptage et les contrôles d'accès granulaires protège les data sensibles.

Les stratégies pour simplifier la gestion de votre data lake

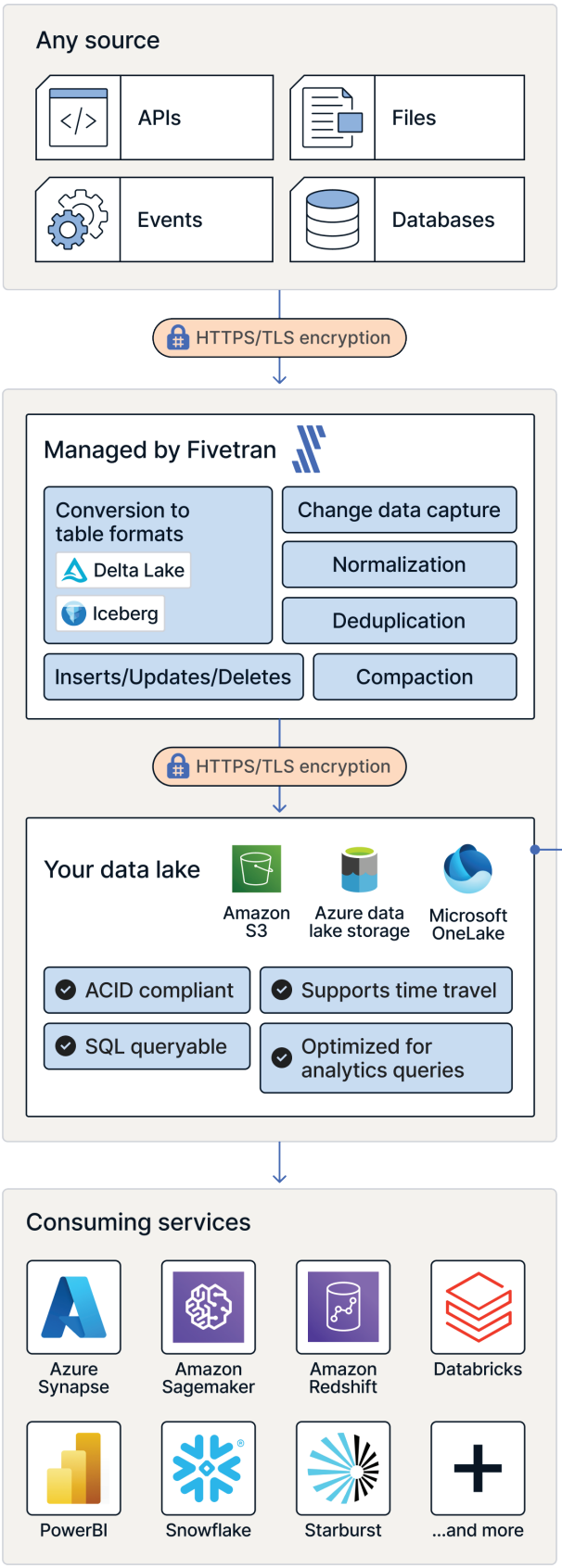
Le service Managed Data Lake (Gestion des data lakes) de Fivetran

En étant capable de transférer plus de 600 sources de data pré-configurées ou personnalisées vers n'importe quelle destination de data lake, Fivetran normalise, compacte et déduplique automatiquement vos data avant de les normaliser dans le format de table Iceberg ou Delta Lake. Fivetran automatise la gestion de votre data lake, depuis le déplacement sécurisé et fiable de data dans des data lakes jusqu'à la maintenance et la mise à jour permanentes des tables.

Le service Managed Data Lake (Gestion des data lakes) de Fivetran :

- **s'adapte intelligemment** aux changements dans vos data,
- **évolue automatiquement** en fonction des modifications de vos schémas,
- **optimise de manière stratégique** vos data afin qu'elles soient propres, cataloguées et normalisées dans un format de table prêt pour l'analyse, et
- **réduit les coûts d'ingestion** en utilisant le service géré de Fivetran pour couvrir les coûts de calcul découlant de l'ingestion de data dans le data lake.

Avant même que vos data n'arrivent dans le data lake, Fivetran fournit de solides contrôles de qualité et de confidentialité des data. Les clients peuvent exclure ou anonymiser les champs PII avant l'ingestion, tandis que Fivetran nettoie, déduplique et normalise automatiquement les data en fonction des standards analytiques. Fivetran écrit ensuite les data nettoyées sur un stockage cloud comme Amazon S3, OneLake et ADLS dans les formats de table ouverts Delta Lake et Iceberg, en renseignant les metadata dans des catalogues de data comme AWS Glue, Polaris et Unity, ce qui permet une préparation immédiate des requêtes. De plus, Fivetran prend en charge les coûts d'ingestion, ce qui permet de libérer des ressources pour les projets d'IA.



Les avantages du service Managed Data Lake (Gestion de data lakes) de Fivetran sont notamment les suivants :

- **Donner aux utilisateurs professionnels et aux spécialistes des data les moyens d'agir** grâce à des data centralisées, démocratisées et prêtes à être interrogées, qui ajoutent du contexte, invitent les informations et favorisent la découverte de data.
- **Augmenter l'efficacité opérationnelle** en convertissant automatiquement vos data en formats de table ouverts (Delta Lake/Apache Iceberg) avec des fonctions robustes de catalogage et de gouvernance des data.
- **Optimiser le temps des développeurs :** Fivetran se charge du gros travail de mise à jour des tables, de déduplication et d'autres tâches de maintenance de bas niveau, ce qui évite aux développeurs de perdre du temps sur des tâches qui peuvent être automatisées.
- **Économiser de l'argent** en automatisant la migration des data depuis des data warehouses coûteux, qui vous enferment dans des formats de data propriétaires, vers des data lakes et des formats de data ouverts, ainsi qu'en couvrant les coûts d'ingestion des data.
- **Fournir une tranquillité d'esprit** grâce à une réplication de data sans souci qui garantit que vos dataset arrivent toujours propres et complets, chaque changement étant capturé.

Gestion manuelle des data lakes	Avantages du service Managed Data Lake (Gestion des data lakes) de Fivetran
Temps important et travail intensif pour les ingénieurs data	Configuration « Set-and-forget », facile pour les ingénieurs data
Complexité et fragmentation des pipelines et des intégrations sur mesure	Sources de data automatisées et gérées à l'aide d'identifiants simples
Mauvaise qualité et incohérences des data, sujettes aux erreurs	Qualité et cohérence des data intégrées
Problèmes d'évolutivité liés à l'augmentation des volumes de data, entraînant des problèmes de pipeline	Évolution sans maintenance à mesure que les volumes de data augmentent
Risques de conformité et de sécurité liés à l'absence de gouvernance et de protection des data	Conformité et sécurité assurées par une gouvernance solide et une protection des data (masquage, hachage)
Frais de maintenance liés aux modifications des sources de data	Gestion automatisée des schémas et fonction de connecteur source de data
Manque de capacités en temps réel	Ingestion de data en temps quasi-réel avec synchronisation toutes les 5 minutes
Manque de compétences en matière de conception, de mise en œuvre et de livraison des pipelines	Interface facile à utiliser et accessible à tous
Augmentation de la dette technique	Service géré sans coûts liés à la dette technique

Fivetran répond à tous vos besoins pour disposer d'un data lake bien gouverné et prêt pour l'analyse :

- **Connectivité complète :** Plus de 600 connecteurs pré-configurés.
- **Multi-plateforme cloud :** Prise en charge d'AWS, Azure et GCP.
- **Sortie au format Lakehouse :** Conversion des data en Iceberg ou Delta Lake, avec les fichiers Parquet sous-jacents.
- **Gestion et maintenance permanente des tables :** Nettoyage en transit, normalisation, évolution du schéma, insertions/mises à jour/suppressions, compactage des fichiers, optimisation des performances.
- **Qualité des data :** Déduplication, évolution et gestion des schémas.
- **Intégration de la gouvernance :** AWS Glue, Databricks Unity Catalog.
- **Rentable :** Frais d'ingestion pris en charge par Fivetran.

Selon une étude IDC, les clients de Fivetran ont réalisé en moyenne 1,5 million de dollars de bénéfices annuels et un ROI de 459 % sur trois ans - des économies que vous pouvez réinvestir pour tirer davantage de valeur de votre data lake.

Source : "The Business Value of Fivetran" par IDC

TÉMOIGNAGE CLIENT

Coupa accélère l'adoption de son data lake S3 avec Fivetran

Défi commercial

Coupa a été confronté à des défis liés à l'ingestion et à la centralisation de data provenant de multiples applications et plateformes dans le cloud. Les silos de data ont entravé sa capacité à générer rapidement des informations sur l'utilisation de la plateforme et le comportement des clients, ce qui a entraîné des opportunités manquées et des informations obsolètes. Coupa a reconnu la nécessité de consolider les sources de data afin d'obtenir une vision globale des interactions avec les clients et des habitudes d'achat, ce qui l'a poussé à rechercher une solution pour briser ces silos de data et permettre des analyses plus efficaces.

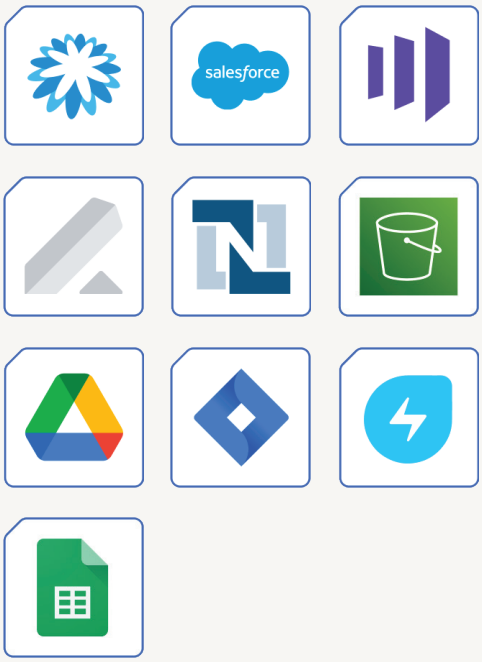
Résultats

Fivetran a permis à Coupa d'automatiser le déplacement de data vers un data lake Amazon S3, réduisant ainsi le délai de rentabilisation de 3-6 mois à seulement quelques semaines. Coupa a triplé son volume de data et ses intégrations, quadruplé le nombre d'utilisateurs de data et amélioré ses capacités de prise de décision stratégique.

Résultats clés :

- Multiplication par 3 du volume de data et des intégrations
- Multiplication par 4 du nombre d'utilisateurs de data au sein de l'organisation
- Amélioration significative de la qualité et de l'exactitude des data
- Réduction du délai de rentabilisation de 6 mois à quelques semaines
- Optimisation de la connaissance client orientée data

Connecteurs



Destination et cloud

Avec le service Managed Data Lake (Gestion de data lakes) de Fivetran, nous rendons les data aussi accessibles et fiables que l'électricité, donnant ainsi aux entreprises les moyens de saisir de nouvelles opportunités et de stimuler l'innovation.

S'inscrire pour un essai gratuit de 14 jours de Fivetran

Fivetran

Fivetran, leader mondial du déplacement de data, aide les clients à exploiter leurs data en vue d'optimiser toutes leurs activités, des applications d'IA et des modèles de ML à l'analyse prédictive et aux charges de travail opérationnelles. La plateforme Fivetran centralise de manière fiable et sécurisée les data de centaines d'applications SaaS et de bases de données vers n'importe quelle destination cloud, qu'elle soit déployée au niveau local, dans le cloud ou dans un environnement hybride. Des milliers d'entreprises internationales, dont Autodesk, Condé Nast, JetBlue et Morgan Stanley, font confiance à Fivetran pour déplacer leurs data les plus précieuses afin d'alimenter l'analyse, de stimuler l'efficacité opérationnelle et de favoriser l'innovation.

Pour en savoir plus, rendez-vous sur Fivetran.com.

Commencez votre essai gratuit

©2024 Fivetran Inc.