



5 MEILLEURES PRATIQUES DE DÉVELOPPEMENT D'UN ENTREPÔT DE DONNÉES

TABLE DES MATIÈRES

- 3** Introduction
- 4** Créer un modèle de données
- 7** Adopter une méthodologie Agile relative aux entrepôts de données
- 8** Préférer l'approche ELT à l'approche ETL
- 9** Utiliser un outil d'automatisation
- 10** Former le personnel aux nouvelles approches
- 11** Résumé
- 12** À propos de Snowflake



INTRODUCTION

La technologie du cloud a révolutionné la façon dont les entreprises accèdent aux données, les stockent et les analysent. Que votre entreprise crée une toute nouvelle plateforme de données ou remanie un système d'entrepôt de données existant pour profiter de nouvelles fonctionnalités, quelques directives et meilleures pratiques assureront la réussite de votre projet. Certaines de ces meilleures pratiques peuvent sembler évidentes, mais les entreprises ne prennent souvent pas le temps de définir et documenter ces points de décision en amont, entraînant ainsi complications et inefficacité.

Dans cet eBook, nous présentons cinq recommandations à suivre pour structurer votre stratégie de données et en assurer la cohérence au sein de votre entreprise, afin que l'entrepôt de données que vous créez réponde aux besoins actuels et futurs de votre entreprise. Ces meilleures pratiques de développement permettront non seulement à toutes les parties prenantes de votre entreprise de valoriser l'entrepôt de données que vous créez, mais aussi d'établir les bases nécessaires pour qu'une plateforme de données d'entreprise au sens large puisse évoluer et s'adapter à l'évolution des besoins de votre entreprise.

1. CRÉER UN MODÈLE DE DONNÉES

Dans tout programme de données, la première étape clé consiste à créer un modèle de données, à savoir une représentation abstraite qui organise les éléments de données et décrit leurs liens entre eux et avec les propriétés de leurs entités réelles. Un modèle de données établit une compréhension et une définition communes des informations importantes pour l'entreprise, ainsi que l'environnement de données global de l'entreprise. Disposer d'un modèle de données vous permet de documenter les ensembles de données qui seront intégrés à l'entrepôt de données, la relation qui existe entre ces ensembles de données et les exigences métier que la plateforme cherche à satisfaire.

Pourriez-vous créer un entrepôt de données sans modèle de données ? Oui, mais en choisissant d'ignorer cette étape de base, vous perdez de précieuses informations. Créer un modèle de données complet est un exercice révélateur pour les entreprises. En effet, différentes équipes fonctionnelles doivent se mettre d'accord sur la définition et la délimitation des ressources de données et exigences métier de l'entrepôt de données avant d'en commencer le développement.

Un modèle de données bien défini a un impact positif longtemps après le lancement de l'entrepôt de données (ou datamart). Par exemple, un modèle de données établit la traçabilité des données pour tous les objets figurant dans l'entrepôt de données, facilitant ainsi l'intégration de nouveaux membres d'équipe et l'ajout de nouveaux objets de données à mesure que les besoins de l'entreprise évoluent.

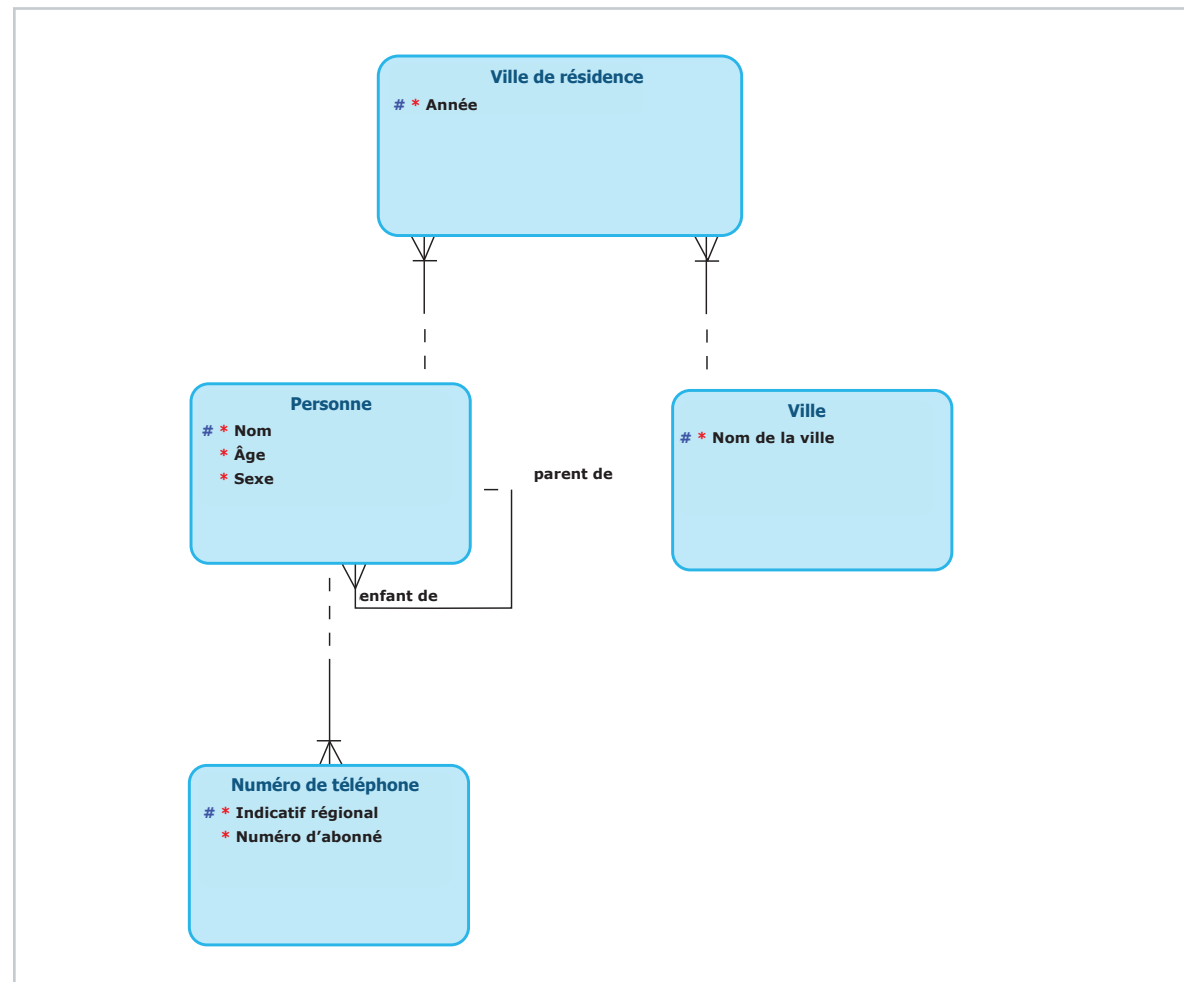


Figure 1 : Modèle de données logique 3NF (troisième forme normale) classique

Le modèle de données permet également une documentation claire du contenu, du contexte et des sources. Il est ainsi plus facile d'effectuer des audits et de respecter les nouvelles exigences relatives aux données, telles que celles édictées par le RGPD (Règlement général sur la protection des données de l'UE, qui définit les directives de collecte et de traitement des informations personnelles).

Un modèle de données bien conçu permet également d'éviter toute confusion et tout remaniement coûteux en aval. Il est toujours judicieux d'intégrer une couche d'intégration indépendante de la source qui permet d'effectuer une analyse sur plusieurs ensembles de données en fonction de leurs points communs.

Un entrepôt de données rassemble différents types et sources de données, notamment des ensembles de données classiques (données de gestion de la relation client [CRM] et données de planification des ressources d'entreprise [ERP], par exemple), mais aussi des blogs, des fils Twitter, des données IDo et même des ensembles de données qui n'ont pas encore été inventés. C'est pourquoi disposer d'une couche d'intégration flexible qui n'est pas trop étroitement liée à un seul système contribue à pérenniser votre entrepôt de données.

Un modèle de données efficace doit utiliser les définitions et structures sémantiques définies par le domaine d'activité, et non les définitions spécifiques d'un seul système source. Par exemple, les clients peuvent être désignés par la mention « cust » dans un système CRM ou par la mention « cust_ID » dans un autre système. Établir une règle sémantique à l'échelle de l'entreprise pour définir la façon dont les utilisateurs peuvent accéder aux données des ensembles de données, les nommer et les analyser est essentiel à la réussite de l'entrepôt de données.

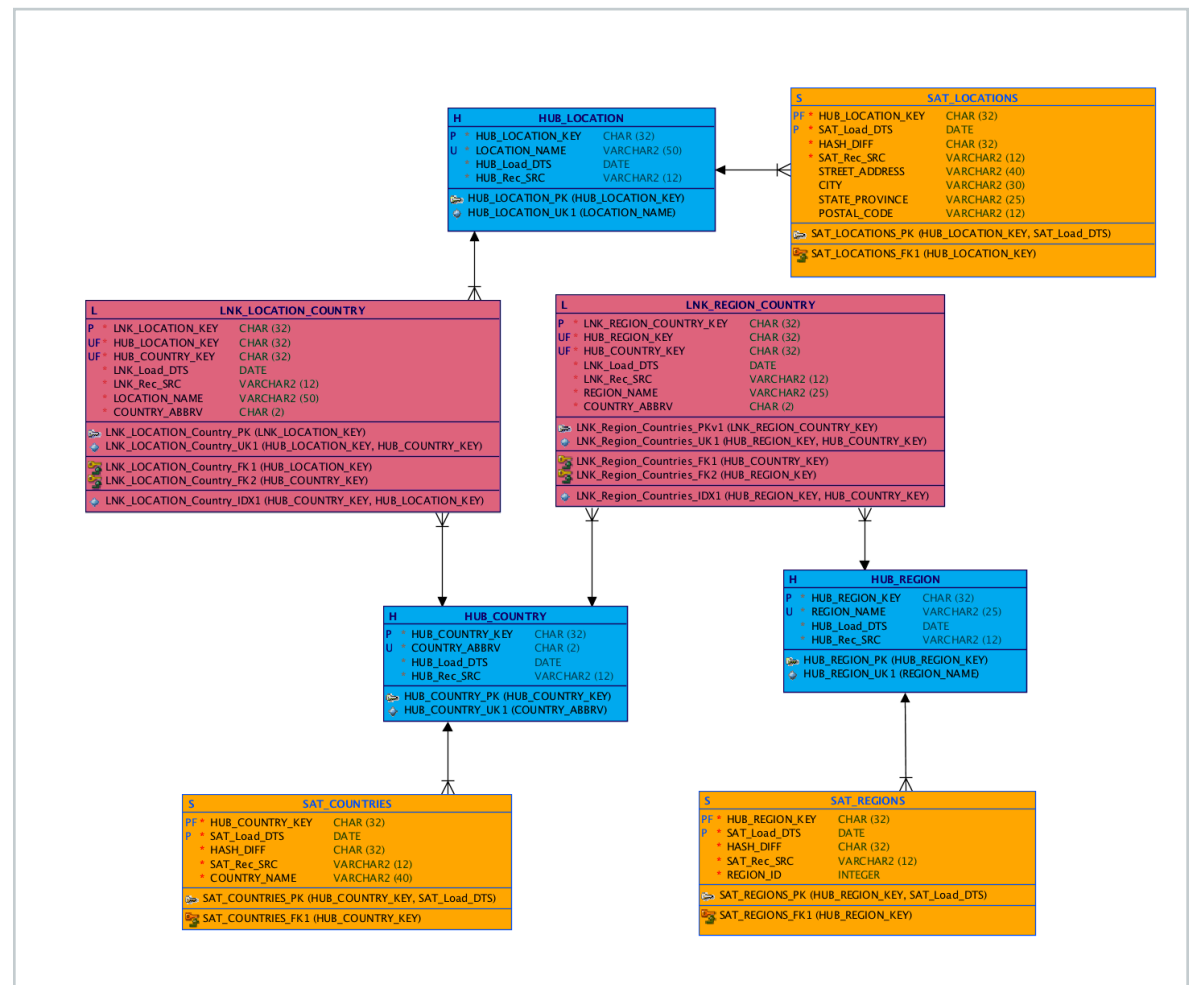


Figure 2 : Exemple de modèle de données utilisant l'approche de modélisation Data Vault

À mesure qu'une entreprise connaît des changements, des fusions ou acquisitions, le système CRM qu'elle utilise aujourd'hui peut être remplacé par un autre. Si votre modèle de données est étroitement lié à un système source spécifique, vous devrez le remanier pour intégrer le deuxième système source qui remplace le système existant. Une couche d'intégration indépendante de la source facilite le mappage des données. Vous pouvez ainsi remplacer un ancien système source par un nouveau sans affecter les rapports en aval et sans avoir à changer le comportement des utilisateurs.

Dans le modèle de données, il est important de choisir une approche standard. Les principales normes de modélisation des données utilisées dans la conception d'un entrepôt de données sont les suivantes :

- **3NF** : 3NF, qui signifie « troisième forme normale », est une norme architecturale conçue pour réduire la duplication des données et assurer l'intégrité référentielle de la base de données.¹
- **Schéma en étoile** : architecture la plus simple et la plus utilisée pour développer des entrepôts de données et des datamarts dimensionnels, le schéma en étoile se compose d'un ou de plusieurs tableaux de faits faisant référence à plusieurs tableaux de dimensions.²
- **Data Vault (DV)** : développée spécifiquement pour résoudre les problèmes d'agilité, de flexibilité et d'évolutivité rencontrés dans les autres approches, la modélisation DV a été créée sous forme d'un référentiel historique, facilement extensible, vérifiable, non volatile et granulaire de données d'entreprise. Cette norme est hautement normalisée et associe des éléments des modèles 3NF et en étoile.³

Chaque architecture a ses avantages, mais votre choix dépendra des besoins de votre entreprise.

Au-delà du type d'architecture que votre entreprise choisit, le plus important est qu'elle choisisse, documente et prenne en permanence en charge cette architecture dans le cadre du développement d'un modèle de donnée pour l'entrepôt de données. Cette approche vous permettra de gagner en efficacité à l'avenir, grâce à une méthodologie unique d'assistance et de dépannage qui permettra aux nouveaux membres de l'équipe d'être opérationnels plus rapidement.



2. ADOPTER UNE MÉTHODOLOGIE AGILE RELATIVE AUX ENTREPÔTS DE DONNÉES

Dans le passé, créer un entrepôt de données (ou même un datamart) demandait d'importants efforts monolithiques sur plusieurs trimestres ou années, le tout soumis au traditionnel processus « en cascade ». De nos jours, ce n'est plus la norme. De nombreuses entreprises choisissent d'adopter une approche de conception plus flexible et itérative, ou Agile.

Face à la constante évolution des besoins des entreprises et à la mise en ligne toujours plus rapide de nouvelles sources de données, les entreprises doivent être capables de s'adapter et d'exploiter ces informations de manière concise et rapide. Il faut donc apprendre à créer des solutions de données et d'analyse de manière agile et incrémentielle. Avec une planification appropriée et adaptée à une couche d'intégration unique indépendante de la source, il est désormais possible de diviser des projets de données importants en composants plus petits qui peuvent être livrés plus fréquemment, apportant ainsi une valeur ajoutée accrue à l'entreprise beaucoup plus rapidement.

Pour atteindre leur objectif, les architectes d'entrepôts de données adoptent la méthodologie Agile, qui a fait sa toute première apparition dans le monde du développement de logiciels. Dans cette méthodologie, les exigences et solutions évoluent par le biais d'une collaboration entre les clients et des équipes autonomes et interfonctionnelles. Lorsqu'elle est appliquée à la conception et à la création d'entrepôts de données, elle permet aux entreprises d'activer de nouveaux ensembles de données et de relever plus rapidement tout nouveau défi de l'entreprise.⁴

Dans le cadre de la méthodologie Agile, diverses approches ont vu le jour pour créer de la valeur plus rapidement :

- **Scrum** : tenant son nom de la phase de jeu de rugby dans laquelle les avants entrecroisent leurs bras et avancent, l'approche Scrum est l'environnement de processus le plus souvent utilisé dans le cadre du développement Agile. La méthode simple Scrum met l'accent sur la communication quotidienne et la réévaluation flexible de plans réalisés lors de phases de travail itératives et courtes.⁵ Ralph Hughes a codifié l'application de la méthode Scrum à l'entreposage de données dans une série d'ouvrages précurseurs utiles aux entreprises adoptant cette approche.
- **Kanban** : la méthode Kanban consiste à gérer la création de produits pour en assurer la continuité sans surcharger l'équipe de développement. À l'instar de l'approche Scrum, la méthode Kanban est un processus conçu pour aider les équipes à collaborer plus efficacement. Tenant son nom des cartes « Kanban » utilisées pour suivre la production dans une usine, la méthode Kanban a été créée par Taiichi Ohno, ingénieur industriel chez Toyota, afin d'améliorer l'efficacité du secteur de la fabrication.
- **BEAM** : la méthode BEAM (Business Event Analysis and Modelling) a été introduite par Lawrence Corr et Jim Stagnitto dans leur ouvrage révolutionnaire « Agile Data Warehouse Design ». La méthode BEAM est axée sur les événements métier, plutôt que sur les exigences de reporting connues, afin de modéliser l'ensemble des processus opérationnels. Elle s'articule autour de sept dimensions (qui, quoi, quand, où, comment, combien et pourquoi) pour identifier et préciser les événements métier.⁶

Pour exploiter pleinement les avantages du développement Agile, une plateforme de données Agile est très utile. Les plateformes de données dans le cloud fournissent une élasticité et flexibilité structurelles, permettant ainsi une évolutivité rapide à mesure que les besoins de l'entreprise évoluent. Les plateformes de données dans le cloud nécessitent moins d'efforts, de maintenance et de tâches administratives pour être efficaces. Elles peuvent évoluer et s'adapter à l'évolution des exigences de l'entreprise. En utilisant un service cloud moderne, les équipes peuvent passer moins de temps à affiner les requêtes et à gérer le stockage, et plus de temps à relever les défis immédiats de l'entreprise et à créer de la valeur commerciale.

L'utilisation de méthodologies et de structures Agile n'est pas une mince affaire. Elle nécessite un engagement culturel au sein de l'entreprise et représente souvent un changement important de mentalité et du flux de travail par rapport aux flux d'entreposage de données traditionnels. Réorganiser une équipe informatique pour travailler confortablement dans un environnement Agile peut prendre de 6 à 12 mois, ce qui peut sembler paradoxal étant donné que la méthodologie Agile a pour objectif de créer de la valeur plus rapidement. Il est possible d'accélérer cette transition en faisant appel à un coach Agile chevronné. Une fois la transition faite, les équipes peuvent commencer à apporter de nouveaux changements incrémentiels à l'entrepôt de données en quelques semaines seulement, au lieu de plusieurs mois.

3. PRÉFÉRER L'APPROCHE ELT À L'APPROCHE ETL

Auparavant, le développement d'un entrepôt de données utilisait une approche ETL (Extract, Transform, Load : extraction, transformation, chargement), consistant à extraire les données à importer dans l'entrepôt de données depuis les systèmes sources, à les nettoyer ou à leur appliquer des règles métier sur un serveur externe, puis à les charger dans l'entrepôt de données cible. Avec l'évolution des fonctionnalités et de la puissance de calcul des plateformes de données, l'approche ELT (Extract, Load, Transform) a vu le jour.

Dans le cadre de l'approche ELT, les données brutes sont extraites de la source, puis chargées, sans grandes modifications, dans la zone de stockage intermédiaire de l'entrepôt de données. Les métadonnées, la date de chargement ou les informations sur la source peuvent être ajoutées aux données, puis être intégrées directement à l'entrepôt de données. Une fois dans l'entrepôt de données, les entreprises peuvent utiliser la puissance de la base de données pour effectuer des transformations : modification de la structure des données (application d'un modèle de données), application de règles métier ou exécution de mesures de qualité des données pour nettoyer les données (correction des adresses incomplètes, standardisation des noms de champs de données et résolution des doublons, par exemple).

L'approche ELT présente deux avantages distincts : économies et traçabilité. L'approche ELT permet de réaliser des économies, car les entreprises peuvent exploiter la puissance de la plateforme de données pour transformer les données, au lieu d'utiliser un serveur externe. Il est généralement beaucoup moins coûteux d'utiliser la puissance de calcul basée sur le cloud que d'exécuter des transformations et de traiter les données sur un serveur externe. Il est donc plus rapide et plus économique de transférer les données directement sur le cloud. L'approche ELT facilite également l'audit et le suivi des données à l'avenir, car elle fournit une image des données sources d'origine directement dans la plateforme de données. Ainsi, l'entrepôt de données lui-même peut jouer le rôle de ce qu'on appelle un « data lake », où les données brutes sont stockées en permanence.



4. UTILISER UN OUTIL D'AUTOMATISATION

L'objectif d'un entrepôt de données est d'activer et de fournir des données plus rapidement afin de permettre aux entreprises de prendre des décisions éclairées et d'augmenter leur valeur. Adopter une méthodologie Agile est un moyen d'y parvenir. Mais il est également possible d'utiliser des outils d'automatisation qui permettent de développer et de déployer un code plus rapidement. Étant donné que de nombreuses méthodologies de développement d'entrepôts de données sont basées sur des schémas, le codage requis pour charger et structurer les données est souvent reproductible, ce qui signifie qu'il peut être automatisé. De nombreux outils du marché automatisent une partie ou la totalité des tâches de conception et de création, et la liste ne fait que s'allonger de jour en jour.

L'automatisation permet aux entreprises d'exploiter pleinement leurs ressources, de répéter les opérations plus rapidement et d'appliquer plus facilement les normes de codage. Elle permet la création d'un code standardisé, ce qui est extrêmement utile dans les entreprises où le code ETL et les modèles de données étaient traditionnellement développés manuellement. L'automatisation fournit une norme documentée pour ces différents artefacts, ainsi qu'un mécanisme d'assurance qualité et d'application pour veiller à ce que tous les développeurs et concepteurs respectent cette norme.

Les outils d'automatisation qui utilisent des modèles pour générer un code sont particulièrement utiles, car ils appliquent des normes en faisant d'elles les propriétés privilégiées au sein des modèles eux-mêmes. Cela accélère l'intégration, car les nouveaux développeurs et concepteurs utiliseront ces outils standard, garantissant ainsi une mise en œuvre uniforme et une courbe d'apprentissage plus courte. Une mise en œuvre uniforme permet de faciliter les tests et débogages, car le code est développé à l'aide des mêmes normes.

Ces outils accélèrent également l'itération, car les générateurs de code automatisés ont tendance à ne pas faire d'erreurs de syntaxe. La mise à jour du code implique généralement l'ajout d'un nouvel objet à l'outil ou la modification des propriétés des modèles au niveau global, générant ainsi un nouveau code immédiatement disponible pour tout déploiement dans l'environnement à des fins de test et de validation.

5. FORMER LE PERSONNEL AUX NOUVELLES APPROCHES

Opter pour la méthodologie Agile ou le développement d'un code automatisé requiert non seulement l'évolution des compétences, mais aussi de votre mentalité. L'équipe doit être formée et entraînée pour tirer parti de ces nouvelles approches et technologies de manière efficace. Cela peut consister à faire appel à des experts externes pour former les équipes aux meilleures pratiques Scrum ou à apprendre aux équipes les avantages, les règles et les meilleures pratiques de l'architecture standard adoptée par l'entreprise pour sa plateforme de données.

De nombreuses ressources industrielles sont disponibles pour aider à gérer la transition vers la méthodologie Agile. **Agile Alliance**, une organisation à but non lucratif mondiale qui s'attache à promouvoir les concepts de développement Agile des logiciels comme décrit dans le manifeste Agile, propose de nombreuses options de formation pour découvrir les concepts Agile. **Scrum Alliance** offre des certifications et des cours de formation de base et avancés pour la méthode Scrum. De même, certains partenaires proposent la certification et le programme de formation Data Vault via **Data Vault Alliance**.

Comme pour tout nouveau processus ou changement culturel, les entreprises doivent gérer la courbe d'adoption pour garantir une transition cohérente et efficace vers la nouvelle approche dans les opérations quotidiennes. Identifier des projets pilotes ou de faisabilité pour initier les équipes aux nouvelles approches permettra aux professionnels de développer et maîtriser les compétences dans des scénarios protégés, mais réels, accélérant ainsi l'apprentissage de ces nouvelles compétences.



RÉSUMÉ

Toutes les meilleures pratiques décrites dans cet eBook requièrent un investissement initial pour atteindre la valeur commerciale à long terme attendue. Mais le retour sur investissement s'articule autour de deux axes : ces meilleures pratiques permettront d'établir les bases d'un programme d'analyse de données efficace dès le départ et d'accélérer la création d'une valeur commerciale incrémentielle pour votre environnement de données et ce, bien après la première mise en production.

Au vu de l'évolution des exigences métier et du souhait de valoriser toujours plus les données et types de données, mettre en place ces meilleures pratiques vous permettra de réfléchir et d'étendre les cas d'usage classiques des entrepôts de données. Avec des bases solides et une plateforme agile, vous serez en mesure d'étendre votre activité à de nouveaux domaines de données et de répondre aux nouvelles demandes en élargissant le programme de sorte à prendre en charge la data science, le machine learning, l'IA et peut-être même la monétisation des données. Avec les ressources cloud flexibles et évolutives actuelles, vos possibilités en matière de données sont sans limites.





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](https://www.snowflake.com)



© 2023 Snowflake Inc. Tous droits réservés. Snowflake, le logo Snowflake et tous les autres noms de produits, de fonctionnalités et de services Snowflake mentionnés dans le présent document sont des marques déposées ou des marques commerciales de Snowflake Inc. aux États-Unis et dans d'autres pays. Tous les autres noms de marque ou logos mentionnés ou utilisés dans le présent document le sont uniquement à des fins d'identification et peuvent être des marques de commerce de leur(s) détenteur(s) respectif(s). Snowflake ne peut être associé à, ou être sponsorisé ou approuvé par, un tel détenteur.

CITATIONS

¹ [fr.wikipedia.org/wiki/Forme_normale_\(bases_de_donn%C3%A9es_relationnelles\)](https://fr.wikipedia.org/wiki/Forme_normale_(bases_de_donn%C3%A9es_relationnelles))

² [fr.wikipedia.org/wiki/%C3%89toile_\(mod%C3%A8le_de_donn%C3%A9es\)](https://fr.wikipedia.org/wiki/%C3%89toile_(mod%C3%A8le_de_donn%C3%A9es))

³ snowflake.com/blog/data-vault-modeling-and-snowflake

⁴ [Agiledata.org/essays/dataWarehousingBestPractices.html](https://agiledata.org/essays/dataWarehousingBestPractices.html)

⁵ scrum.org/resources/what-is-scrum

⁶ bisystembuilders.com/beam