



COMMENT TISSER VOTRE DATA MESH SUR SNOWFLAKE

Au cours des dernières années, une approche de la gestion des données appelée « data mesh¹ » a su s'imposer. Des entreprises issues de tous les secteurs choisissent le data mesh pour la gestion décentralisée de leurs données. Cette solution a été pensée pour favoriser l'agilité des données et éviter les goulots d'étranglement organisationnels trop souvent liés à des approches centralisées et monolithiques.

Ce livre blanc traite de l'approche du data mesh adoptée par Snowflake. Il présente les fonctionnalités Snowflake les plus utiles au data mesh et évoque les différentes options d'architecture traditionnelles adoptées par nos clients lors de la mise en œuvre d'une plateforme de données en libre-service prenant en charge les domaines distribués.

Le data mesh est avant tout une approche organisationnelle destinée à attribuer différentes responsabilités à chaque équipe chargée des domaines et à leurs produits de données respectifs, tout en les coordonnant. Mais pour que le concept de data mesh soit applicable aux domaines, vous devez impérativement vous doter de la bonne technologie.

« **Le data mesh n'est pas une question de technologie [...], mais il suppose une technologie adaptée et riche en fonctionnalités qui se met au service des équipes chargées des produits de données. Cette solution ne demande pas aux équipes chargées des domaines et des produits de données de réinventer la roue et de construire de toute pièce leur propre plateforme de données et d'analyse. Nous devons simplifier la tâche des équipes chargées des produits de données. C'est indispensable aux notions de responsabilisation et de décentralisation.** »

— OMAR KHAWAJA, Global Head BI chez Roche (2022)

De nombreuses entreprises adeptes du data mesh utilisent Snowflake comme plateforme de données et les résultats sont positifs. Il n'existe pas encore une seule et unique plateforme technologique offrant une solution complète de bout en bout pour prendre en charge le concept de data mesh. Cependant, Snowflake dispose de nombreuses fonctionnalités essentielles au fonctionnement d'une plateforme de données en libre-service. Cette solution est notamment dotée d'une architecture distribuée fondée sur les domaines, mais aussi de fonctionnalités destinées à favoriser la transformation de données en produit et la mise en œuvre de la gouvernance de calcul fédérée.

L'APPROCHE DU DATA MESH ADOPTÉE PAR SNOWFLAKE

Snowflake a accompagné les initiatives data mesh de nombreux clients et tire de ces expériences l'approche suivante :

- Le data mesh représente avant tout une transformation organisationnelle. Bien qu'elle nécessite souvent des changements au niveau de l'architecture informatique et de la technologie employée, cette transformation a de nombreuses implications non techniques.
- Rester pragmatiques. Lorsque nous conseillons nos clients, l'objectif n'est pas le data mesh « idéal », mais de résoudre leurs problématiques. Par exemple, le stockage polyglotte et l'accès multimodal sont des concepts utiles, mais les entreprises doivent axer leur stratégie sur leurs exigences réelles pour davantage d'impact.
- Ne pas viser la lune dès le début. Commencez modestement et déployez votre stratégie data mesh petit à petit en laissant mûrir votre approche. Commencez par répondre à un besoin immédiat de votre entreprise en l'appliquant à un ou deux domaines/produits de données. Plus tard, tirez parti du succès rencontré pour étendre la portée de votre data mesh.
- Maîtriser les coûts et la complexité. Par exemple, il a été prouvé qu'avoir une plateforme de données en libre-service dotée d'un ensemble d'outils aussi petit et cohérent que possible a de nombreux avantages, et ce dans tous les domaines. Cette configuration permet aussi de répondre à toutes les exigences essentielles de chaque domaine.
- Définir les incitations et les critères de réussite au plus tôt. Ces éléments incluent notamment la mise en place d'indicateurs clés de performance (KPI) mesurables pour les domaines, les produits de données, la plateforme de données en libre-service et les contrôles de gouvernance.
- Il n'existe aucune solution de data mesh prête à l'emploi. Pour créer des solutions adaptées aux exigences de nos clients, nous pouvons compter sur notre vaste réseau de partenaires. Bien qu'ils ne soient pas détaillés dans cet article, les outils utilisés pour la gouvernance des données, l'automatisation, les DevOps et bien d'autres domaines sont souvent indissociables d'une solution data mesh.

¹ www.thoughtworks.com/en-us/what-we-do/data-and-ai/data-mesh

LES FONCTIONNALITÉS PERTINENTES PROPOSÉES PAR SNOWFLAKE

Snowflake offre un certain nombre de fonctionnalités que nos clients jugent essentielles à la construction d'une plateforme de données en libre-service adaptée au data mesh.

Snowflake est bien plus qu'un entrepôt de données dans le cloud

Snowflake est un fournisseur de services cloud intégrés doté d'un large éventail de fonctionnalités destinées au data engineering, aux data lakes, au data warehousing, au data sharing et à d'autres aspects importants du cycle de vie traditionnel du machine learning.

Les utilisateurs peuvent notamment créer et automatiser des pipelines de transformation des données pour convertir diverses données saisies en produits de données gouvernés. Snowflake est aussi à l'aise dans le traitement de formats de fichiers courants de vos compartiments de stockage dans le cloud que dans celui de flux d'entrées (depuis Kafka par exemple) ou de tables relationnelles.

Les formats de fichiers pris en charge incluent les formats suivants : JSON, XML, Parquet, AVRO, Delta Lake², Apache Iceberg³ et d'autres. Snowflake prend également en charge les données non structurées telles que les images, les vidéos et tout autre type de format binaire. Les données peuvent être traitées dans la plateforme Snowflake à l'aide de SQL, Python⁴, Scala, Java et JavaScript. Le traitement peut aussi être effectué en appelant les fonctions externes sur la plateforme de cloud dans son ensemble.

Snowflake peut être le fournisseur idéal de fonctionnalités pratiques dont ont besoin vos équipes chargées des domaines, mais il peut aussi ne pas vous correspondre. Il n'en demeure pas moins vrai que nous vous offrons une large gamme de fonctionnalités regroupées au sein d'une seule et même plateforme. Avec d'autres solutions, vous devriez intégrer tout un ensemble de services cloud. L'intégration de ce type de services peut être complexe, chronophage et gourmande en main-d'œuvre qualifiée.



FIGURE 1 : SNOWFLAKE, UNE PLATEFORME UNIQUE CONÇUE POUR DIFFÉRENTS TYPES DE DONNÉES ET DIFFÉRENTS WORKLOADS

² En public preview au moment de la publication (août 2022).

³ En private preview au moment de la publication (août 2022).

⁴ En public preview au moment de la publication (août 2022).

La plateforme distribuée Snowflake n'a rien de monolithique

Snowflake est une plateforme distribuée et interconnectée conçue pour éviter les silos, mais aussi pour permettre aux équipes distribuées d'échanger des données de manière contrôlée et sécurisée. Comment fonctionne la plateforme ? Une entreprise peut créer un ou plusieurs comptes Snowflake rattachés à la même région cloud/plateforme ou depuis une région cloud/plateforme différente (Figure 2). Chaque compte peut accueillir plusieurs bases de données séparées. Les ressources de calcul et de stockage pour ces dernières peuvent être déployées et mises à l'échelle de manière indépendante et distribuée.

Grâce à la puissance de calcul indépendante des bases de données ou des comptes, les différentes équipes chargées des domaines peuvent travailler de manière autonome, tout en utilisant la plateforme Snowflake sous-jacente pour partager des ressources de données. Il faut noter que Snowflake ne conçoit pas les bases de données au sens traditionnel et relationnel du terme. Pour nous, une base de données inclut également toutes nos fonctionnalités pratiques (data engineering, data lake, data warehousing, data sharing et data science). La plateforme Snowflake est dotée d'une fonctionnalité essentielle qui exploite des clusters de calcul pour combiner et traiter les données provenant de plusieurs bases de données ou de plusieurs comptes.

Snowflake intègre les fonctionnalités de data sharing et de Marketplace Snowflake

Dans Snowflake, les producteurs de données ont la possibilité de partager des données, des services de données ou des applications avec d'autres comptes

en publiant des métadonnées (« listes »). Grâce aux contrôles de découverte de listes, les producteurs ont la possibilité de partager des informations avec d'autres comptes ou groupes de comptes en toute confidentialité, ou publiquement via la Marketplace Snowflake. Les données partagées par les producteurs de données peuvent être accompagnées de SLA ou de SLO. Ces éléments incluent la fréquence de mise à jour des données, leur historique, leur granularité temporelle et toute autre propriété destinée à décrire les données comme un produit.

D'autres équipes peuvent lancer des recherches pour découvrir les ressources de données pertinentes mises à leur disposition, et en obtenir ou demander l'accès. Ces utilisateurs de données bénéficient d'un accès en direct aux données partagées. Cependant, ces dernières restent sous le contrôle du producteur, qui se réserve le droit de personnaliser les politiques d'accès ou de révoquer l'accès aux données à tout moment. Pour accéder aux données partagées, le producteur et l'utilisateur n'ont pas besoin de mettre en œuvre un processus d'ETL ou de mouvement des données. Les producteurs peuvent également publier et partager des tables externes qui permettent d'avoir une « vue » des fichiers stockés en dehors de Snowflake. Ces tables offrent la possibilité d'inclure les formats Delta Lake et Iceberg. Les producteurs peuvent même partager des données avec des tiers qui ne font pas partie de leur entreprise, et ce même si ces tiers ne sont pas des clients actifs de Snowflake. Un producteur de données est donc en mesure de partager des données en externe grâce à un compte de lecteur Snowflake, mais aussi grâce à l'ensemble des API prises en charge. Il peut aussi choisir d'exporter périodiquement des données (partitionnées) aux formats de fichiers courants vers un compartiment de stockage dans le cloud.

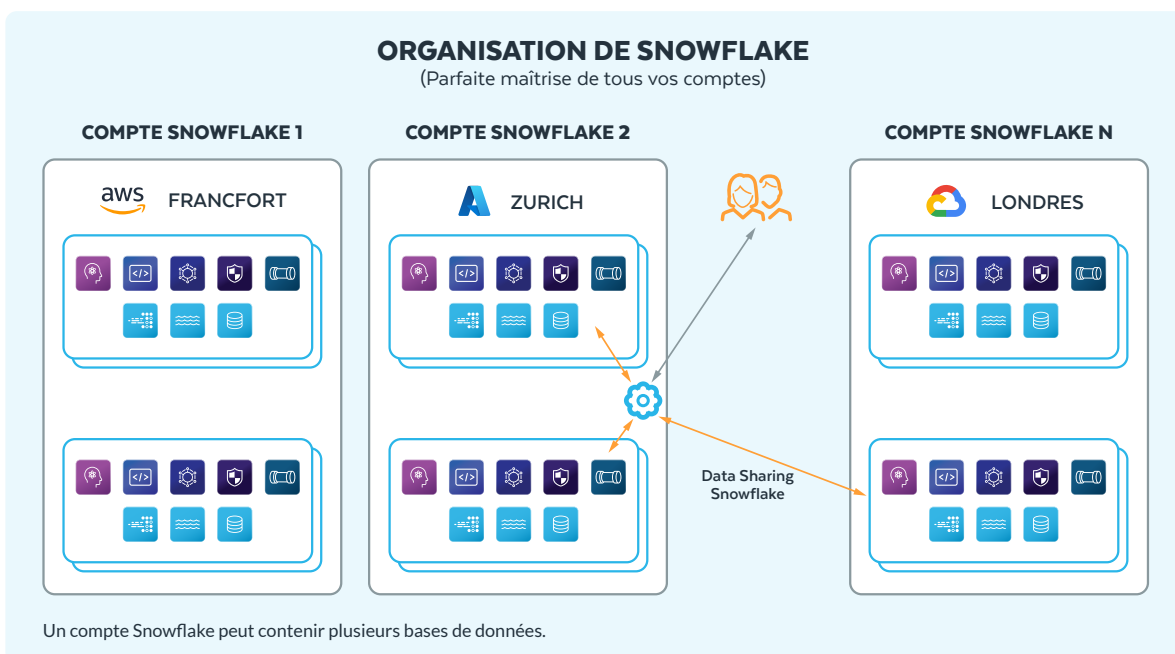


FIGURE 2 : L'ORGANISATION, LES COMPTES ET LES BASES DE DONNÉES SNOWFLAKE PRENNENT EN CHARGE UNE ARCHITECTURE DISTRIBUÉE

Snowflake vous offre un large éventail de fonctionnalités dédiées à la sécurité et à la gouvernance

Dans le processus de déploiement du data mesh, la gouvernance fédérée est sans doute l'un des aspects les plus difficiles à mettre en œuvre puisqu'il nécessite souvent la combinaison d'un ou plusieurs outils pour pouvoir répondre à toutes les exigences. La plateforme Snowflake prend en charge le contrôle d'accès basé sur les rôles, les politiques d'accès au niveau des lignes, le masquage des données au niveau des colonnes, la tokénisation externe, la traçabilité des données, les fonctionnalités d'audit et plus encore. Dans Snowflake, les utilisateurs auront aussi la possibilité d'associer une ou plusieurs balises de métadonnées (paires clé-valeur) à la plupart des objets suivants : comptes, bases de données, schémas, tables, colonnes, clusters de calcul, utilisateurs, rôles, tâches, partages, etc. Les balises sont héritées de la hiérarchie d'objets. Vous pouvez donc les exploiter pour découvrir, suivre, restreindre, surveiller et auditer des objets en fonction de la sémantique définie par l'utilisateur. En outre, les politiques d'accès basées sur les balises⁵ sont idéales pour les utilisateurs souhaitant associer une restriction d'accès à une balise. Ainsi, la politique d'accès est automatiquement appliquée à tout objet de données correspondant qui porte la balise indiquée.

Dans Snowflake, la plupart des contrôles de gouvernance (balises, politiques d'accès, règles de masquage, etc.) peuvent être définis indépendamment de l'application de ces contrôles aux objets de données. Ainsi, les propriétaires de domaine peuvent s'accorder sur les balises ou stratégies courantes utilisées dans les différents domaines, tout en laissant chaque domaine appliquer ou étendre ces éléments de manière individuelle. En outre, les vues sécurisées et la fonctionnalité de data clean room se mettent au service de l'analyse des données sensibles qui ne peuvent pas être partagées d'une autre manière.

Des données concernant l'utilisation d'un produit ainsi que des indicateurs tels que la télémétrie et les données de consommation sont collectés pour pouvoir être utilisés en cas d'analyse de l'impact. Grâce à cette collecte, les équipes chargées des domaines peuvent observer la façon dont les différents utilisateurs utilisent leurs produits de données, tout en analysant leur fréquence d'utilisation.

Snowflake vous offre une expérience apparentée à du libre-service

Généralement, nos clients choisissent Snowflake pour sa facilité d'utilisation et pour la maintenance quasi nulle que nécessite la plateforme. Pour nous, il s'agit de caractéristiques essentielles pour une plateforme en libre-service. Les utilisateurs n'ont pas besoin d'être assistés par l'équipe chargée des infrastructures informatiques : ils peuvent instancier et mettre à l'échelle leurs propres clusters de calcul en un clin d'œil. Le clonage des environnements de développement et de test est tout aussi intuitif. Vous pouvez configurer un mécanisme de saisie des données de changement grâce à 1 ligne d'instruction DDL (Data Definition Language) SQL. Toutes les fonctionnalités de la plateforme Snowflake ont été conçues pour être intuitives.

⁵ En public preview au moment de la publication (août 2022).

PRODUITS DE DONNÉES DANS SNOWFLAKE

Dans un data mesh, chaque domaine crée, gère et possède un ou plusieurs produits de données qu'il partage avec d'autres domaines et utilisateurs de données. Pour envisager des données comme un produit, il faut avant tout adopter une mentalité orientée produit et en faire une habitude organisationnelle. En outre, les domaines ont besoin d'outils en libre-service adaptés qui prennent en charge la création et la gestion des produits de données. Découvrez comment Snowflake peut vous aider à transformer vos données en produit.

Par définition, un produit de données est une combinaison de données, de métadonnées, de code et de dépendances infrastructurelles.

- **Données** : dans Snowflake, les données d'un produit de données se présentent sous forme de tables, de vues, de fichiers (JSON, XML, Parquet, Avro, CSV, etc.) ou de tables externes qui vous donnent un aperçu des fichiers disponibles en dehors de Snowflake. Un seul et même produit de données peut être constitué de plusieurs de ces objets. Généralement, les domaines regroupent les objets de données grâce à l'utilisation d'un seul et unique schéma par produit de données. Parfois, ils exploitent aussi le code de chaque produit de données. Pour répondre aux besoins des utilisateurs de données, les producteurs de données peuvent modéliser les données de la manière qu'ils jugent la plus appropriée.
- **Métadonnées** : les métadonnées d'un produit de données incluent les métadonnées techniques de ses objets de données, telles que les noms de tables ou de colonnes, les types de données ou les définitions des formats de fichiers. Les métadonnées incluent également les dépendances d'objet, la traçabilité des données et l'historique des accès. Des balises (paires clé-valeur) peuvent être associées à chaque objet pour exposer des métadonnées arbitraires telles que l'origine des données, le domaine, la sensibilité, les termes économiques, la taxonomie, le centre de coûts et d'autres attributs définis par l'utilisateur.

Lorsqu'un producteur publie un produit de données sur la Marketplace Snowflake, il est invité à fournir de la documentation contenant les éléments suivants : description du produit, besoins de l'entreprise, exemples, conditions de service et lien vers le service d'assistance associé au produit de données. Le producteur est également invité à mentionner les SLO des produits de données. Ces éléments incluent la fréquence de mise à jour des données, leur historique, leur granularité temporelle et toute autre propriété (voir Figure 3).

- **Code** : le code d'un produit de données est constitué de pipelines et de transformations destinés à créer et à rafraîchir un produit de données. Dans Snowflake, ce code peut inclure des tâches Snowflake, des canaux, des flux, des procédures stockées⁶, des fonctions définies par l'utilisateur, etc. Tous ces objets Snowflake peuvent être regroupés par produit de données dans un schéma. Le code de ces objets utilise les langages SQL, Java, JavaScript, Scala ou Python et s'exécute nativement sur la plateforme Snowflake.

Le code peut aussi inclure des politiques. Dans Snowflake, ce code peut servir au contrôle d'accès basé sur les rôles, aux politiques de masquage dynamique des données, aux politiques de contrôle d'accès au niveau des lignes, aux vues sécurisées, au balisage d'objet, mais aussi au classement ou à l'anonymisation/tokénisation des données.

- **Dépendances infrastructurelles** : une tâche Snowflake destinée à planifier et à orchestrer le pipeline pour rafraîchir un produit de données peut être associée à un certain cluster de calcul. Il peut s'agir d'une ressource de calcul dédiée à un seul produit de données ou partagée entre plusieurs produits de données. Dans tous les cas, le cluster peut être suspendu et repris de manière automatique dès que vous en avez besoin. Ainsi, vous évitez de générer des coûts lorsque le cluster n'est pas en cours d'exécution. Les clusters peuvent aussi être adaptés à la hausse ou à la baisse comme dans un outil en libre-service. Pour réduire ou éliminer les dépendances infrastructurelles explicites, les tâches, canaux et autres opérations peuvent s'exécuter sans serveur.

FIGURE 3 : INDICATION DES SLO D'UN PRODUIT DE DONNÉES POUR UNE LISTE DE PRODUITS DE DONNÉES

⁶ En public preview au moment de la publication (août 2022).

Snowflake prend en charge divers ports d'entrée et de sortie pour les produits de données, dont l'ingestion de flux, un connecteur Kafka, un connecteur Spark, une API Dataframe, l'ingestion des données automatique depuis les compartiments de stockage dans le cloud, une API REST, les formats de fichiers, mais aussi les API SQL telles que JDBC, ODBC, .NET et les API destinées aux langages de programmation les plus courants. Axées sur la sécurité et la transparence, les fonctionnalités

collaboratives de la plateforme Snowflake vous offrent la possibilité d'accéder aux données, aux services de données et aux applications, mais aussi de les distribuer dans plusieurs clouds sans pipeline ETL et sans intégrations.

Les produits de données présentent, eux aussi, un certain nombre de propriétés importantes. Le Tableau 1 vous donne des exemples de fonctionnalités Snowflake conçues pour vous aider à bénéficier de ces caractéristiques.

CARACTÉRISTIQUES D'UN PRODUIT DE DONNÉES	LISTE DES FONCTIONNALITÉS DE SNOWFLAKE (NON EXHAUSTIVE)
Sécurisé	Contrôle d'accès basé sur les rôles, politiques d'accès au niveau des lignes, masquage dynamique des données, chiffrement et tokenisation
Visible	Découverte ciblée/Marketplace Snowflake et intégration optionnelle d'un catalogue tiers
Adressable	Partage des données Snowflake et accès standardisé cross-cloud et cross-régions
Intuitif	Balises de métadonnées personnalisées, listes de données avec documentation et forme statistique des données dans Snowsight
Fiable	SLA/SLO tels que la fréquence de mise à jour des données ou leur granularité, la traçabilité des données, les dépendances d'objets ou encore l'historique d'accès
Accessible nativement	SQL, Python, Java, Scala, API SQL, API REST, Dataframes et autres éléments de programmation pour accéder à des données multimodèles (structurées, semi-structurées, non structurées, avec différents types de fichiers, etc.)
Interopérable	Types de données conformes à la norme ANSI SQL, métadonnées unifiées, API communes à tous les domaines, collaboration, data sharing Snowflake, Marketplace et data exchange Snowflake
Autosuffisant	Produits de données mixtes composés de plusieurs objets, les produits de données peuvent être constitués d'objets de données et de fonctions pouvant être partagés avec les utilisateurs de produits de données

TABEAU 1 : CARACTÉRISTIQUES DES PRODUITS DE DONNÉES PRISES EN CHARGE DANS SNOWFLAKE

OPTIONS D'ARCHITECTURE POUR LES DOMAINES DISTRIBUÉS

Revenons plus en détail sur les différentes topologies Snowflake adoptées par les entreprises à la recherche d'une plateforme de prise en charge des domaines distribués. Ces topologies servent de modèles généraux, mais dans les faits, leur mise en œuvre peut dépendre des exigences et des préférences spécifiques.

- « **Compte par domaine** » : chaque domaine utilise un compte Snowflake séparé.
 - Isolation maximale entre les domaines.
 - Différents domaines peuvent fonctionner dans différentes régions et plateformes de cloud.
 - Cette option permet un data mesh multi-régions et multi-cloud avec une expérience Snowflake cohérente. Elle offre également des fonctionnalités intégrées de data sharing entre les domaines basées sur un système d'échange de données et de métadonnées central grâce auquel tous les domaines peuvent publier des produits de données et y accéder.
- « **Base de données par domaine** » : chaque domaine utilise une ou plusieurs bases de données Snowflake séparées.
 - Toutes ces bases de données sont gérées depuis un seul compte Snowflake.
 - La gestion des utilisateurs, de la sécurité et de la gouvernance de tous les domaines est simplifiée.
 - L'accès aux produits de données est facilité par la définition d'autorisations au niveau de l'objet pour les différentes bases de données.
 - Chaque équipe chargée des domaines peut encore déployer et faire évoluer ses propres clusters de calcul indépendamment de ceux des autres domaines.
- « **Schéma par domaine** » : chaque domaine utilise des schémas séparés au sein d'une base de données unique.
 - Degré d'isolement le plus faible entre les environnements de domaine.
 - Chaque équipe chargée des domaines peut encore déployer et mettre à l'échelle ses propres clusters de calcul sans pour autant les inclure à d'autres domaines.
 - Efforts potentiellement plus importants dans les conventions d'appellation destinées à distinguer les objets des différents domaines.
 - Utile pour les sous-domaines dans un scénario domaine/sous-domaine.

- Dans le reste de cet article, nous ne détaillerons pas l'option « Schéma par domaine », mais elle présente de nombreuses similitudes avec l'option « Base de données par domaine ».
- « **Domaines hétérogènes** » : les domaines peuvent utiliser différentes piles informatiques.
 - Certains domaines utilisent Snowflake et d'autres domaines utilisent des systèmes tiers.
 - Certains domaines sont basés sur le cloud et d'autres sont on-premise.
 - Niveau de complexité plus élevé pour s'adapter aux environnements de domaine hétérogènes.
 - Cette option peut aller à l'encontre de l'objectif du data mesh qui consiste en l'utilisation d'une plateforme en libre-service commune et indépendante du domaine applicable à tous les domaines. Nous vous invitons à faire preuve de vigilance.

Des variations d'architecture ou des approches hybrides dérivées de ces options de base sont susceptibles d'émerger. Par exemple, une entreprise peut choisir l'option « Base de données par domaine » tout en gérant ses bases de données avec plusieurs comptes. Certains domaines pourraient aussi utiliser des bases de données séparées là où d'autres exploitent l'intégralité de leur compte Snowflake. Généralement, l'environnement dans lequel évolue l'équipe chargée des domaines se compose d'outils additionnels en fonction de leurs exigences et compétences, qui viennent compléter les fonctionnalités de Snowflake.

Ce qu'il faut retenir, c'est que Snowflake s'adapte à de nombreux choix d'architecture pour permettre différents compromis entre, d'une part, autonomie du domaine et décentralisation, puis d'autre part, degrés de complexité différents et gestion opérationnelle.

Chaque entreprise doit trouver le juste équilibre entre centralisation et décentralisation. La solution adoptée doit être adaptée à sa taille, à son historique et à sa culture organisationnelle. Cette recherche d'équilibre s'applique aussi à la gouvernance fédérée. Les entreprises doivent trouver une solution entre contrôle centralisé et autonomie locale du domaine.

Les sections suivantes détaillent davantage ces options d'architecture. Ici, nous allons délaisser l'intégration d'outils tiers que nos clients utilisent souvent en complément de Snowflake ou de leurs autres initiatives data mesh pour nous focaliser sur les topologies Snowflake.

Compte unique : base de données par domaine

Généralement, nos clients adeptes du data mesh adoptent une topologie populaire qui consiste en l'utilisation d'un compte Snowflake unique. Au sein de ce dernier, les domaines utilisent des bases de données et des clusters de calcul séparés comme environnements autonomes. Pour répondre à leurs besoins en matière de développement, de test et de production, les différents domaines peuvent être associés à un(e) ou plusieurs bases de données/ clusters. Plateforme en libre-service, Snowflake permet aux domaines d'utiliser sa fonctionnalité « Clonage zéro copie » pour (re)créer des environnements de développement et de test de manière instantanée et à la fréquence désirée. En outre, les différents utilisateurs d'un seul et même domaine peuvent être autorisés à déployer leurs propres clusters de calcul, mais aussi à les adapter à leurs besoins respectifs. Cependant, des moniteurs de coût et de consommation ainsi que des quotas peuvent être configurés en fonction des domaines ou d'autres niveaux de granularité de la hiérarchie des utilisateurs et des ressources.

Le Data Cloud Snowflake permet à tous les domaines d'avoir leurs environnements et leurs ressources de calcul séparés à disposition sans devoir être des silos physiques, ce qui compliquerait l'accès aux produits de données.

La gouvernance est définie de manière centralisée et appliquée à toutes les bases de données dotées d'un processus DevOps. Cette gouvernance peut être simplifiée par la mise en œuvre de fonctionnalités telles que des balises d'objet pour conserver une vue d'ensemble des différents objets appartenant aux domaines. Au sein des domaines, la gouvernance est contrôlée par les équipes chargées des domaines. Pour sécuriser les données et empêcher les utilisateurs ou domaines d'accéder par erreur à des données, ces équipes mettent en place un contrôle d'accès basé sur les rôles ainsi que des politiques d'accès au niveau des lignes et des colonnes.

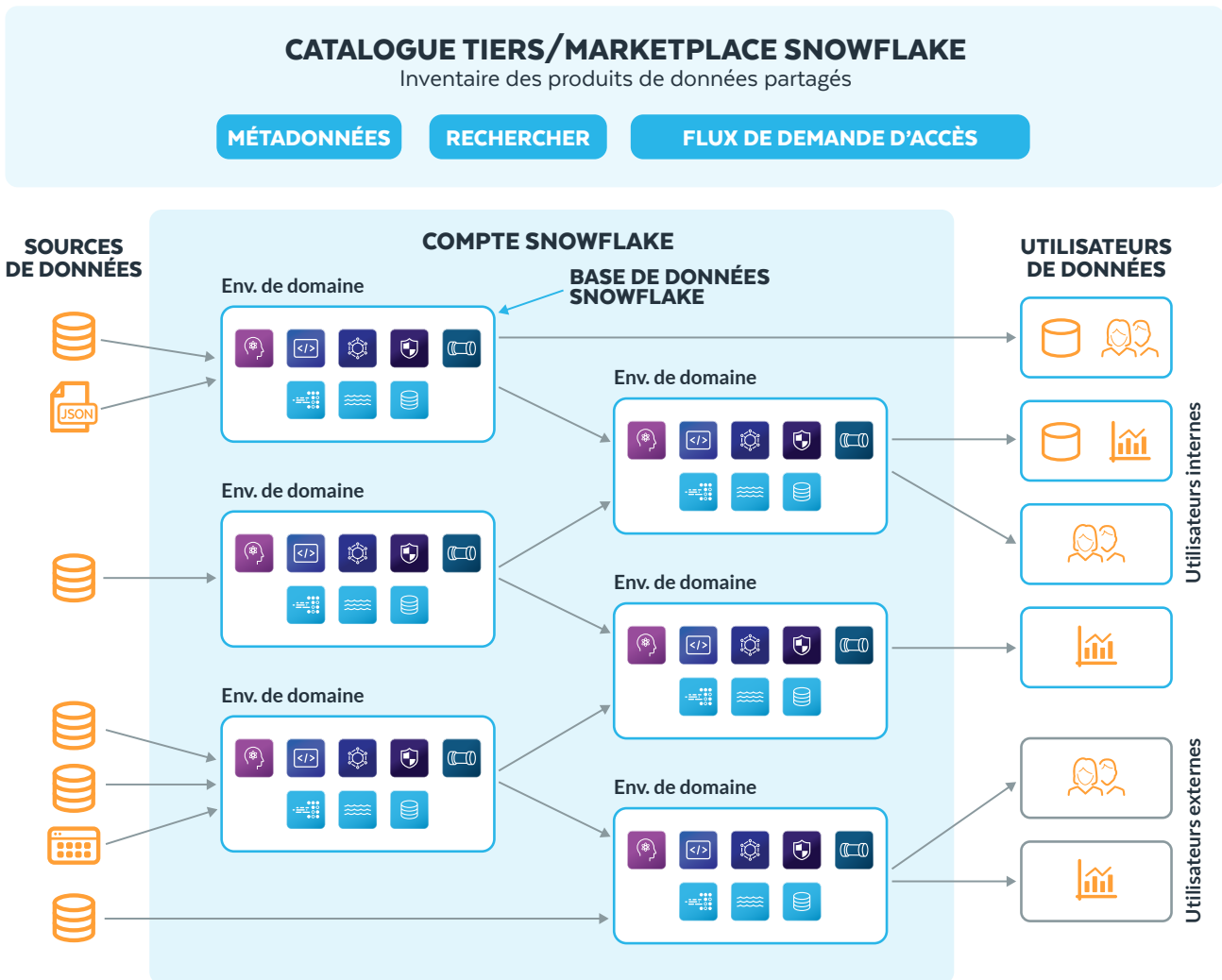


FIGURE 4 : COMPTE SNOWFLAKE UNIQUE – BASE DE DONNÉES PAR DOMAINE

Chaque domaine peut disposer de plusieurs schémas. L'un d'entre eux peut notamment servir de couche destinée à mettre les produits à disposition d'autres domaines. Vous pouvez aussi suivre une autre approche qui consiste en l'utilisation d'une base de données partagée et commune, chaque domaine pouvant disposer d'un schéma pour publier ses produits de données sous forme de vues (pas de copies). En fonction de leurs caractéristiques, ces produits peuvent être constitués de données structurées, semi-structurées ou non structurées. Pour plus de visibilité, les produits sont ensuite répertoriés dans un catalogue de données tiers.

Il existe plusieurs manières de demander l'accès à un produit. Il existe notamment l'approche manuelle, qui invite le demandeur à ouvrir un ticket, envoyé à l'équipe chargée des domaines. Cette dernière le traitera avant de refuser ou d'autoriser l'accès au demandeur qui, le cas échéant, se verra attribuer un rôle lui garantissant l'accès. Certains catalogues proposent un flux plus automatique.

En regroupant tous vos domaines au sein d'un seul et même compte Snowflake, vous bénéficiez des avantages suivants :

- Accès rapide aux produits de données grâce à la définition d'autorisations intra-bases de données.
- Politique d'administration des réseaux, de la sécurité et de la gouvernance centralisée et simplification de la gestion globale par ricochet.
- Simplification de la récupération après sinistre avec l'assistance d'un seul autre compte localisé dans une autre région ou dans un autre cloud.

Étant donné que chaque domaine a potentiellement besoin d'environnements DT(A)P (Développement, Test, Acceptation, Production), qu'il peut facilement créer avec le clonage zéro copie, le nombre d'objets peut être considérable. Les conventions d'appellation doivent donc être planifiées avec prudence.

Cette approche est similaire à l'utilisation de l'option « Schéma par domaine ». Tout est organisé de manière logique au sein d'un seul et même compte Snowflake. Les conséquences sont donc similaires à celles de l'approche « Base de données par domaine ». Pour ce qui est du partage public de données avec des utilisateurs externes via la Marketplace Snowflake ou du partage privé via les contrôles de découverte de listes, il vaut mieux disposer d'une base de données par domaine.

Comptes multiples : compte par domaine

Pour permettre à chaque domaine de fonctionner avec un compte Snowflake distinct, vous pouvez utiliser une autre topologie. Ces comptes peuvent être localisés dans la même région ou plateforme cloud, ou dans des régions et plateformes cloud différentes. Avec sa portée mondiale, le Data Cloud Snowflake permet aux entreprises et aux domaines de partager des données entre les comptes, les régions et les plateformes de cloud. Ces entreprises et domaines disposent également d'un accès standardisé, sécurisé et contrôlé aux produits de données de leurs pairs. Cette fonctionnalité rencontre un franc succès auprès de nos clients qui souhaitent bénéficier d'une solution data mesh multi-régions et multi-cloud.

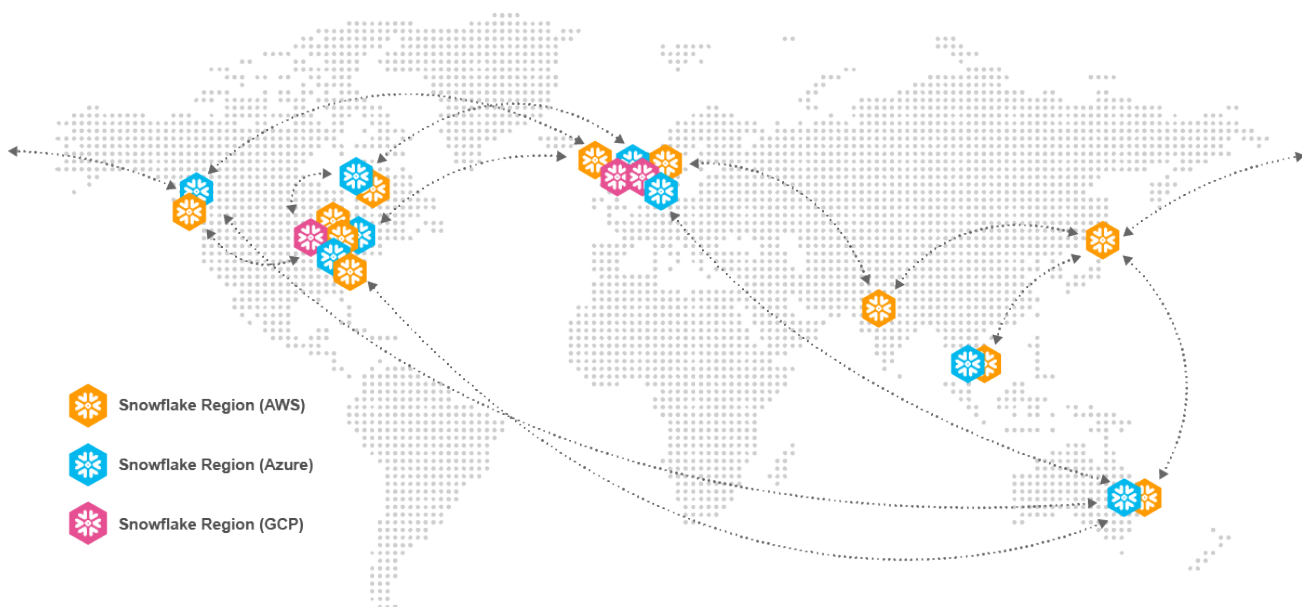


FIGURE 5 : SNOWGRID COMME PLATEFORME DATA CLOUD MONDIALE

Les entreprises choisissent cette topologie pour de nombreuses raisons. Par exemple, les différents domaines d'une entreprise qui opère à l'échelle mondiale et de manière distribuée pourraient s'aligner naturellement aux différents emplacements et aux différentes régions du monde. Certaines entreprises à la présence mondiale sont susceptibles de devoir respecter les exigences liées à la localité des données. C'est notamment le cas des entreprises internationales dont certaines données ont interdiction de quitter l'Europe sans être anonymisées, masquées ou soumises à d'autres mesures destinées à garantir la conformité aux réglementations relatives à la confidentialité des données.

Les fusions et acquisitions sont aussi un bon exemple d'événements courants qui peuvent forcer des entreprises à échanger des données d'une région ou d'une plateforme cloud à une autre. D'autres entreprises, elles, adoptent une stratégie multi-cloud pour se diversifier ou pour s'adapter aux préférences et aux investissements existants réalisés par différentes unités commerciales. Dans les cas où une gestion des utilisateurs et de la sécurité séparée est nécessaire, l'utilisation de comptes multiples permet également une plus grande autonomie des domaines.

Par conséquent, la topologie résultante (Figure 6) est très similaire à celle qui préconise l'utilisation d'une base de données séparée par domaine. La seule exception réside dans la mise à disposition d'un compte Snowflake propre à chaque domaine et dans l'utilisation des fonctionnalités de data sharing et la Marketplace Snowflake pour rendre les produits de données accessibles à d'autres domaines.

Si nous comparons cette approche à l'approche « Base de données par domaine », nous pouvons noter les avantages suivants :

- Les fonctionnalités de data sharing et de collaboration Snowflake peuvent être utilisées dans tous les domaines.
- L'application des normes d'appellation globales est simplifiée par le caractère indépendant de l'espace de noms de chaque compte.
- La plateforme cloud et les préférences régionales peuvent être prises en charge.
- Chaque compte dispose d'une gestion de la sécurité et des utilisateurs qui lui est propre.

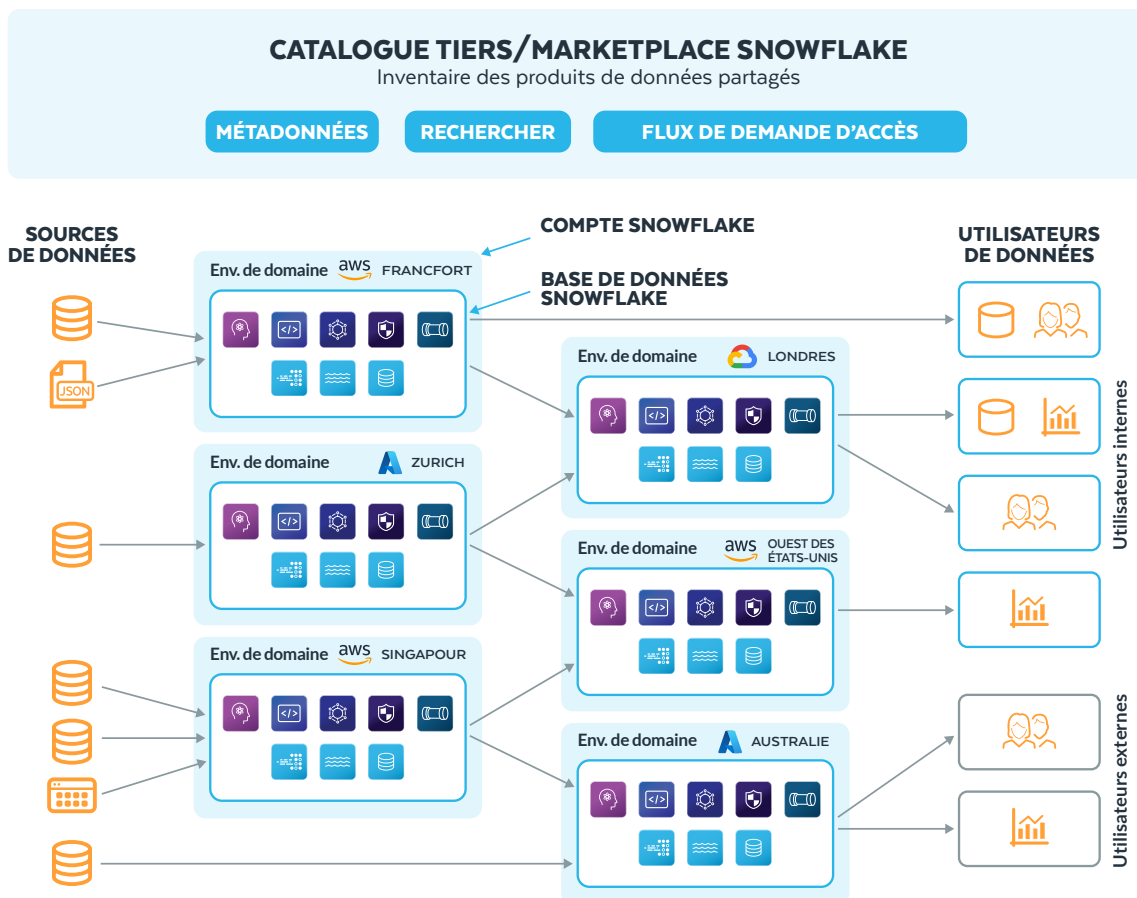


FIGURE 6 : COMPTES MULTIPLES – UN COMPTE PAR DOMAINE

Architecture hétérogène

Certains de nos clients souhaitent savoir comment intégrer des environnements de domaine extérieurs à Snowflake aux topologies décrites ci-dessus. Ces intégrations favorisent la création d'une architecture hétérogène au sein de laquelle les domaines n'utilisent pas tous la même plateforme de données indépendante du domaine pour mettre en œuvre leurs pipelines et leurs produits de données. Souvent, elles sont motivées par le désir de réutiliser plusieurs piles technologiques ou plusieurs référentiels qui existent déjà au sein de l'organisation.

Nous avons pu constater qu'une architecture aussi hétérogène augmente généralement le coût et la complexité du processus de data mesh. Ces conséquences sont intrinsèquement liées à la grande hétérogénéité des systèmes participants, qui ne facilite pas la cohérence de domaines essentiels tels que la gouvernance, la sécurité, les métadonnées, les normes d'interopérabilité, les performances, les compétences requises, l'assistance informatique et plus encore. Nous encourageons donc nos clients à prêter attention au rôle joué par les divers systèmes et référentiels qu'ils cherchent à intégrer au data mesh. Ces systèmes sont-ils réellement utilisés comme environnements de domaine au service des produits de données ? Ou devraient-ils plutôt être

considérés comme des sources de données que les domaines traitent comme une entrée ? Dans ce cas, les clients peuvent revenir aux topologies présentées ci-dessus.

Pour intégrer des domaines extérieurs à Snowflake, nos clients peuvent adopter une approche consistant à déplacer des ressources de données ou des « produits de données quasiment prêts » vers une couche intermédiaire. Snowflake sera ainsi en mesure d'agir comme un « proxy » destiné à servir une gouvernance cohérente, de la sécurité, de l'interopérabilité et bien plus encore au reste des produits de données du data mesh.

Les sujets Kafka intégrés à Snowflake par ingestion continue et par des mises à jour automatiques des produits de données dans Snowflake pourraient faire office de couche intermédiaire. La couche intermédiaire pourrait également se composer d'un ou de plusieurs compartiments de stockage dans le cloud d'Amazon S3, d'Azure Blob Storage, d'Azure Data Lake Storage ou de Google Cloud Storage. La liste des formats de données pris en charge inclut, notamment, les suivants : JSON, XML, Parquet, AVRO, Apache Iceberg, Delta Lake, etc. Deux options s'offrent à Snowflake : l'ingestion automatique

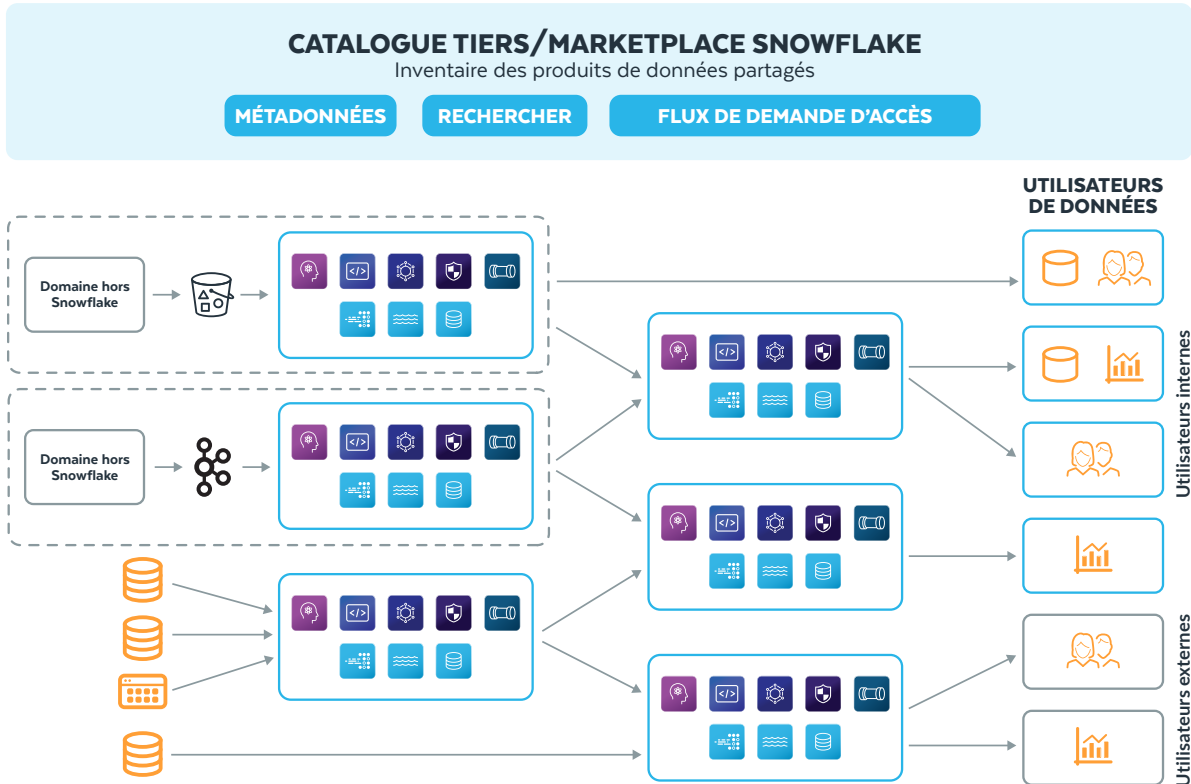


FIGURE 7 : ARCHITECTURE HÉTÉROGÈNE

et continue de nouveaux fichiers provenant de compartiments de stockage pour des performances, une sécurité et une gestion automatique optimales, ou la présentation de l'accès en lecture de fichiers tels que des tables externes au reste du data mesh. Les tables externes de Snowflake consistent en des vues des données disponibles dans des fichiers en dehors de Snowflake. Pourtant, les tables externes de Snowflake sont des objets de données de première classe qui, comme d'autres objets de données Snowflake, peuvent être sécurisés, contrôlés, joints et même partagés via l'outil Collaboration Snowflake. Ces tables hissent Snowflake au rang de couche d'intégration destinée à exposer les données externes de manière cohérente et contrôlée, sans pour autant intégrer et dupliquer lesdites données.

Si vous souhaitez intégrer des environnements extérieurs à Snowflake à la plateforme de données, il existe d'autres solutions. Malheureusement, ces dernières dépassent le cadre de ce document.

Pour certaines entreprises, la virtualisation des données apparaît comme une solution potentielle pouvant contribuer à l'intégration d'un ensemble diversifié d'environnements de domaine. Bien qu'il existe des cas d'usage favorables à la virtualisation des données, nous avons constaté que cette méthode

vient, elle aussi, avec son lot de défis. Lorsque des données provenant de plusieurs référentiels doivent être fusionnées, la performance pose un véritable problème. Généralement, pour surmonter ce défi, il est nécessaire de déplacer les données vers un emplacement commun pour calculer la fusion, même si d'autres prédicats peuvent être poussés vers les sources de données.

Cette action peut gêner la virtualisation des cas d'usage où la performance est un facteur critique. Avec Snowflake, l'action de fusionner plusieurs objets de données en une seule base de données présente des caractéristiques de performances similaires aux situations dans lesquelles ces objets de données se trouvent dans des bases de données ou dans des comptes Snowflake séparés. Cette propriété est à la base même de l'architecture de la plateforme Snowflake. Pour certaines entreprises, la virtualisation est un autre grand défi. On lui reproche notamment d'encourager les équipes à continuer à travailler dans leurs îlots technologiques respectifs, souvent très spécifiques à un domaine, au lieu d'essayer de déployer une plateforme en libre-service commune et indépendante du domaine.

RÉSUMÉ

Le data mesh n'est pas un remède miracle aux divers problèmes de gestion et d'intégration des données de votre entreprise. Mais si vous estimez que le data mesh est l'approche la plus adaptée pour votre entreprise, oubliez les considérations techniques et concentrez-vous sur l'aspect organisationnel indispensable à la pérennité de cette solution. Vous pouvez notamment vous focaliser sur les changements organisationnels, les rôles et responsabilités, la gestion du personnel, les incitations et la responsabilisation, l'adhésion des parties prenantes essentielles ou encore l'évolution de votre mentalité vers une approche orientée produit.

Un jour ou l'autre, vous devrez songer à concevoir une architecture informatique en libre-service capable de prendre en charge des domaines distribués et des produits de données avec gouvernance fédérée. En sa qualité de plateforme en libre-service intuitive destinée aux équipes chargées des domaines, Snowflake peut jouer un rôle déterminant dans cette démarche. Snowflake prend en charge différentes topologies pour aider les entreprises à choisir le niveau de décentralisation et d'autonomie de domaine souhaité, tout en assurant l'interconnectivité et l'interopérabilité des domaines. Grâce à Snowflake, bénéficiez de topologies à compte unique ainsi que d'architectures multi-régions et multi-cloud. La plateforme prend également en charge l'intégration de domaines externes ou de configurations de collaboration multi-entreprises. La plateforme Snowflake sous-jacente, avec Snowgrid et la Marketplace Snowflake, aident ensemble les entreprises à éviter de créer des silos de données.

En outre, Snowflake offre une large gamme de fonctionnalités destinées à accompagner les entreprises dans la transformation de leurs données en produits, mais aussi dans l'application du concept de gouvernance fédérée. En plus de cela, Snowflake s'intègre facilement à une large gamme d'outils tiers pour offrir davantage de fonctionnalités à votre plateforme. Si vous souhaitez accompagner les changements organisationnels et les processus nécessaires à la réussite de la transformation du data mesh, le Data Cloud Snowflake est un allié technologique de choix. Pour en savoir plus sur la façon dont Snowflake peut vous aider, rendez-vous sur snowflake.com/data-mesh

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 octobre 2022, des milliers de clients de nombreux secteurs, dont 543 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).



© 2022 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.