

QUALITÉ ET GOUVERNANCE DES DONNÉES

Optimiser la valeur des données à l'ère de l'IA

Obtenez des résultats plus rapides et une meilleure gouvernance avec les data products

Alors que notre environnement professionnel change rapidement et constamment, les données sont des ressources stratégiques précieuses qui optimisent les opérations, stimulent l'innovation et accélèrent la croissance. Ce ne sont pas juste des chiffres et des statistiques. Elles renferment des insights, des tendances et des modèles qui contribuent à la réussite des entreprises. Portées par l'engouement suscité par l'IA générative et le machine learning, les organisations doivent cependant faire face à une question cruciale :

Comment exploiter toute la valeur des données pour optimiser leurs résultats, tout en renforçant la confiance dans les données et leur gouvernance ?

Tout est question de rapidité. Selon une étude récente menée par IDC, 75 % des décideurs pensent que quelques jours suffisent à faire perdre leur valeur aux données.¹

Le véritable enjeu ne se limite pas à la gestion d'énormes volumes de données. Il s'agit également d'affiner l'information en la transformant en assets d'excellente qualité, soigneusement répertoriés et facilement accessibles, adaptés à des domaines métier spécifiques. Quiconque travaillant avec des données sait que cela est plus facile à dire qu'à faire. En effet, les experts dans les domaines de la donnée ou de l'analytics sont confrontés à plusieurs défis qui les empêchent de concrétiser pleinement leurs initiatives data :

- **La gestion des données dans le cadre de projets n'est pas assez évolutive.** Chaque équipe projet travaille avec ses propres paramètres, notamment en matière de champ d'application, d'objectifs et de parties prenantes. Une telle approche fragmentée rend difficile la collaboration transversale et limite la valorisation des données. Les risques sont multiples : désynchronisation des parties prenantes dans les différents projets, duplication des efforts, écarts de gouvernance, utilisation inefficace des données, augmentation des coûts et, en fin de compte, retards ou objectifs non remplis.

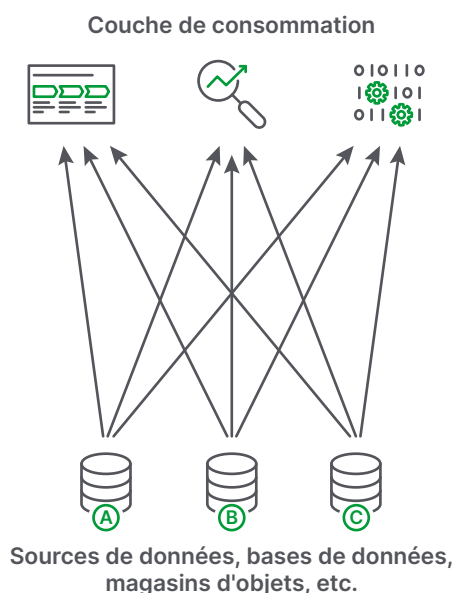
- On constate, à tous les niveaux de l'entreprise, **un manque de cohérence dans la qualité des données à travers les différents systèmes**, ce qui compromet la prise de décision, altère les processus et sape la confiance dans les insights data-driven.
- Tout le long du cycle de vie des données et de la chaîne de valeur, une certaine confusion règne : **qui, des producteurs ou des consommateurs de données, en détient la propriété et en est responsable ?** Cette question entrave les efforts d'extraction de valeur à mesure que l'entreprise se développe.
- **Le recours à des solutions fragmentées et à des approches ad hoc et bricolées** de la gestion des données ralentit la création de valeur et s'avère coûteux lorsqu'il faut les étendre à l'ensemble de l'entreprise et lorsqu'il faut assurer leur maintenance en l'absence d'un personnel qualifié.

Ces enjeux creusent le fossé entre producteurs et consommateurs de données. D'un côté, on a les producteurs de données (focalisés sur l'IT), qui donnent la priorité aux plateformes de données plutôt qu'à la compréhension de l'utilisation des données et des besoins des consommateurs. De l'autre, les consommateurs de données (bien ancrés dans leurs domaines métier) ont du mal à transmettre leurs exigences et ne font pas confiance aux données fournies. **Pour combler cet écart grandissant, les organisations ont besoin d'une approche stratégique et optimisée de la gestion des données.**

¹ IDC Blog, « Navigating the Planes of Enterprise Intelligence Architecture », juin 2023.

Révolutionner la gestion des données

Les entreprises ont généralement recours à deux approches bien distinctes pour extraire de la valeur à partir de leurs données : l'approche **fragmentée** et l'approche **centralisée**.



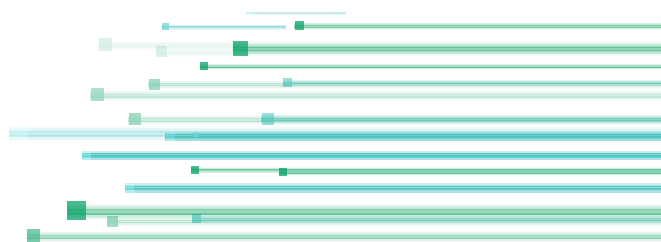
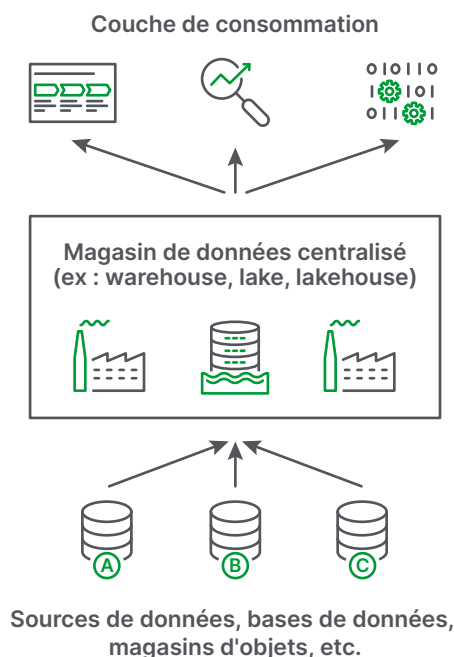
L'APPROCHE FRAGMENTÉE

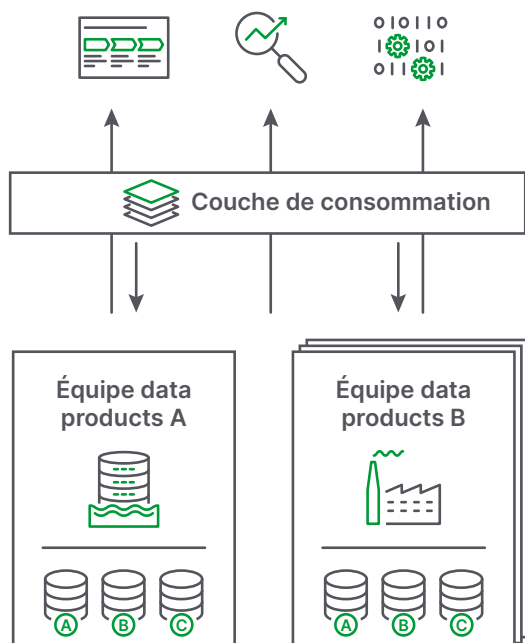
L'approche fragmentée consiste à mettre sur pied plusieurs équipes de « projets de données », chargées de collecter les données provenant de divers systèmes sources pour répondre à des cas d'usage spécifiques de l'entreprise. Chacune d'entre elles se sert de ses propres outils et technologies pour extraire, intégrer et analyser les données en fonction de ses besoins spécifiques. Cette approche permet de générer une valeur immédiate à court terme. Cependant, elle n'est pas viable à long terme car elle entraîne la multiplication des silos de données, la duplication des efforts, l'augmentation des coûts et des retards dans l'accomplissement des objectifs.

L'APPROCHE CENTRALISÉE

L'approche centralisée, quant à elle, consiste à confier à une équipe unique la responsabilité d'extraire, de nettoyer et d'agrégier les données à grande échelle. Celle-ci a pour objectif de répondre de manière exhaustive aux besoins analytiques de l'entreprise. Les membres de cette équipe établissent des pipelines de données complexes pour récupérer les données de divers systèmes, les nettoyer et les consolider dans des data lakes, des warehouses ou des lakehouses dans le cloud (tels que Databricks Delta Lake, Snowflake, BigQuery, S3, etc.).

La centralisation peut sembler résoudre les problèmes liés à la fragmentation. Elle présente cependant un grave défaut : elle nécessite une coordination très poussée entre les départements pour effectuer des changements ou des mises à jour, minant au passage l'agilité opérationnelle. De plus, puisque l'équipe data centrale se concentre sur les exigences communes de l'entreprise dans son ensemble, les préférences et attentes spécifiques des parties prenantes et des consommateurs sont souvent laissées de côté. La solution fournie n'est donc pas optimale et ne s'aligne pas sur les besoins métier.





L'approche fédérée offre plusieurs avantages clés :



Agilité et flexibilité

En rationalisant la gestion des données de bout en bout, une approche fédérée donne la possibilité aux organisations de libérer **rapidement de la valeur** à partir des données. Elle favorise l'agilité en instaurant un état d'esprit axé sur le produit, dans lequel la livraison et la création de valeur sont réparties entre des équipes produits réparties dans différents services, tandis que les fonctions d'ingénierie de la plateforme sont centralisées. Cela permet de **réagir plus rapidement à l'évolution des besoins métier**, tout en garantissant l'intégrité des données et la responsabilité de chacun vis-à-vis de ces dernières tout au long du processus.



Rapport coût-efficacité

Le modèle fédéré intègre la **réutilisation**, ce qui permet aux organisations d'optimiser leurs investissements et de réduire leurs coûts opérationnels. Par ailleurs, cette approche permet de **réduire significativement la redondance de tâches et d'améliorer la maintenabilité à long terme** grâce au partage centralisé des composants fondamentaux entre les différentes équipes produits.

L'APPROCHE FÉDÉRÉE

Les deux approches classiques (fragmentée et centralisée) présentent des limites. Elles ne permettent pas de répondre de façon optimale aux besoins des entreprises en matière de consommation de données. Heureusement, un changement de paradigme est en cours. L'**approche fédérée**, qui tend à se généraliser, consiste à répartir de façon stratégique la livraison de données en exploitant les data products et en gérant, de manière centralisée, les fonctions d'ingénierie de la plateforme.

Ainsi, **il est possible de faire travailler en parallèle plusieurs équipes data products**, chacune d'entre elles se concentrant sur ceux spécifiques à son domaine. Cette approche permet d'accélérer l'obtention de résultats grâce à une collaboration transversale, y compris entre les parties prenantes techniques et métier.



Accélération de la mise en œuvre des cas d'usage

Une collaboration facilitée entre des équipes transversales permet d'obtenir **rapidement des résultats spécifiques à un domaine**, et adaptés à l'évolution des besoins de l'entreprise et aux nouvelles initiatives data-driven. L'approche fédérée favorise **l'évolutivité et la personnalisation de cas d'usage spécifiques à un domaine** par le biais de streams de développement parallèles au sein de plusieurs équipes data products.



Confiance accrue

Le modèle fédéré permet **d'améliorer la gouvernance, la transparence et la conformité**, et renforce la confiance dans la prise de décisions data-driven. Définir clairement qui possède les données favorise l'alignement et la responsabilisation des parties prenantes en matière de data et d'analytics, ce qui renforce encore le cadre de gouvernance des données de l'entreprise.

Data products : de quoi parlons-nous au juste ?

Vouloir mettre en œuvre une approche fédérée constitue le meilleur point de départ pour les data products. En effet, pour de nombreuses organisations, elle se traduit par un passage du contrôle à la collaboration, du projet au produit et de la rigidité à l'agilité, tout en libérant de la valeur à grande échelle. Le **data mesh** et la **data fabric**, deux approches distinctes de la gestion des données, peuvent former ensemble une puissante synergie pour exploiter tout le potentiel des data products. Dans le cas du data mesh, la propriété des données est décentralisée, détenue et gérée par les équipes qui les utilisent et son aspect le plus concret et le plus important est le data product. La data fabric devient alors l'architecture privilégiée pour la gestion des données, permettant de fournir de manière optimale des data products aux équipes métier.²



Selon l'enquête Gartner CDAO Agenda Survey for 2024, 50 % des organisations participantes ont déjà déployé des data products, et 29 % se sont engagées à piloter ou à envisager un déploiement au cours de l'année prochaine.³

Mais au juste, qu'est-ce qu'un data product ? Un data product fournit des datasets de grande qualité, traités, prêts à l'emploi, exploitables par l'IA, facilement accessibles et applicables à différents cas d'usage métier à l'échelle d'une organisation. On peut assimiler un data product à un microservice pour les données. De même que les microservices mettent en exergue les capacités d'une application par le biais d'unités plus petites, plus contextuelles et composables, un data product permet d'accéder à des données pertinentes et spécifiques à un domaine par le biais d'interfaces de services. Un data product respecte un certain nombre de principes fondamentaux :



Il doit être spécifique à un domaine et avoir un propriétaire

Chaque data product doit être adapté à un domaine spécifique et avoir un propriétaire de données (personne ou équipe) responsable de sa performance, de sa valeur et de son adoption.



Il doit être pensé comme un produit

Les data products sont élaborés en appliquant la notion de produit aux données, en ce qui concerne la propriété, le versioning et les clients (ici, les consommateurs de données). Leur amélioration est dictée par la compréhension de leurs modes d'utilisation et le retour d'information des clients. Tout comme les produits réels, ils ont un cycle de vie.



Il doit être normalisé et interopérable

Les data products respectent des formats, des schémas et des conventions de métadonnées normalisés. Ils facilitent l'interopérabilité et l'intégration avec d'autres composants de l'écosystème de données.



Il doit être identifiable

Les data products sont enregistrés dans un catalogue de data products et peuvent être facilement trouvés et utilisés par les consommateurs de données. Chacun d'entre eux se caractérise par la description des données qu'il représente, la manière dont on peut y accéder, sa sémantique métier, son origine et sa qualité.



Il doit être réutilisable

Les data products sont élaborés une seule fois, mais doivent être exploitables à de multiples reprises dans différents cas d'usage. Certains data products sont élaborés à partir de data products existants.

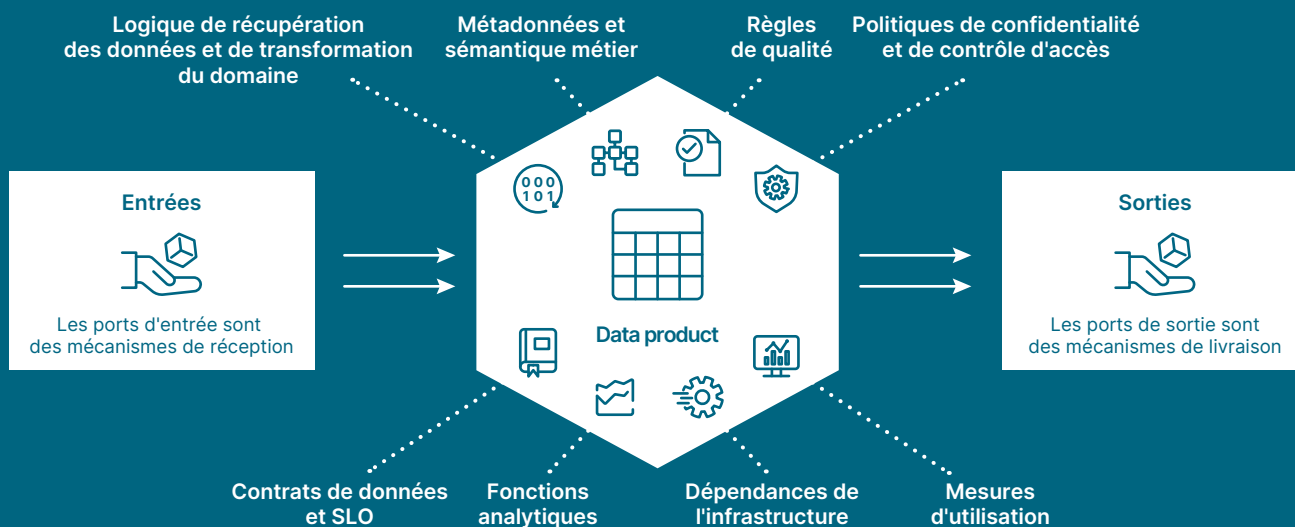


Il doit être accessible

Les data products sont accessibles via un ensemble standard d'interfaces, telles que SQL ou une API REST, qui sont définies et exposées lors de la création des data products. Ces interfaces dissimulent la complexité du data product.

² Gartner: 6 Lessons Data Leaders Can Learn from Adopters of Data Mesh, Robert Thanaraj, et al. 4 mars 2024. GARTNER est une marque déposée et une marque de service de Gartner, Inc. et/ou ses affiliés aux États-Unis et dans le monde. Elle est utilisée ici dans le cadre d'une permission préalable. Tous droits réservés.

³ Gartner: Chief Data and Analytics Officer Agenda Survey for 2024. 12 mars 2024. Bien que les résultats de cette étude ne soient pas représentatifs des conclusions globales ou du marché dans son ensemble, ils reflètent assez bien les impressions des personnes interrogées et des entreprises sondées. GARTNER est une marque déposée et une marque de service de Gartner, Inc. et/ou ses affiliés aux États-Unis et dans le monde. Elle est utilisée ici dans le cadre d'une permission préalable. Tous droits réservés.

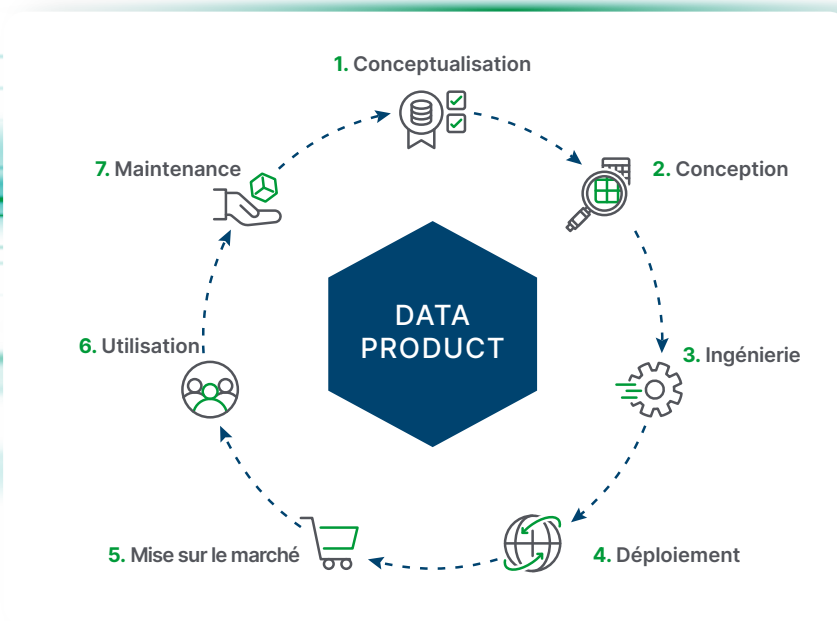


Un data product est un quantum autonome qui regroupe les données, les transformations, le code, les politiques de gouvernance, les accords de niveau de service (SLA), les modèles d'accès et les dépendances de l'infrastructure en une seule unité. Le diagramme ci-dessus décrit la structure d'un data product.

- 1 **Dataset** : se trouve au centre du produit. Il peut s'agir d'un ensemble de tableaux, de fichiers, de graphiques ou de représentations vectorielles. Un data product peut être composé de plusieurs niveaux de données : un ensemble brut qui représente les données telles qu'elles ont été ingérées d'un système source, un ensemble traité qui représente les données nettoyées et intégrées, et un dataset prêt à être consommé, conçu pour faciliter l'utilisation.
- 2 **Entrées** : représentent les ports d'entrée ou les sources de données entrantes. Il peut s'agir de tables dans une application SAP qui doivent être ingérées, de tables dans Snowflake qui doivent être utilisées comme source, d'un autre data product à partir duquel un nouveau produit est dérivé ou d'une file d'attente de messages pour recevoir des flux de données en temps réel.
- 3 **Sorties** : représentent les mécanismes de livraison de données dédiées à la consommation. Les interfaces standardisées d'un data product peuvent être très diverses : SQL, API REST, recherche vectorielle, message et fichier. Les ports de sortie peuvent également varier en fonction du type de service fourni par le data product. Un port de sortie peut donner accès à des données historiques, tandis qu'un autre donnera accès aux données les plus récentes disponibles en temps réel.
- 4 **Logique de récupération des données et de transformation du domaine** : représente la logique et le code du pipeline de données afin d'acquérir les données (ex : migration de données d'une base données source dans une plateforme de données cloud ou lecture d'un ensemble de tables). Il s'agit ensuite d'appliquer une logique de transformation pour façonner les données du format source au format du dataset cible.
- 5 **Métadonnées et sémantique métier** : contient les métadonnées techniques (noms des tables ou des fichiers, définitions des colonnes) et les métadonnées métier (domaine, nom et type du produit, finalité, propriétaire, tag, etc.)
- 6 **Règles de qualité des données** : définissent les différents contrôles de qualité des données qui doivent être exécutés sur le data product. Chaque règle comporte des informations de configuration portant sur le moment de l'exécution, le champ d'application de la règle, l'exécution sur l'ensemble des données ou l'exécution incrémentielle, etc. Les règles de qualité des données seront exécutées suivant un calendrier défini afin de garantir l'exactitude des données et d'instaurer la confiance.
- 7 **Politiques de confidentialité et de contrôle d'accès** : instaurent les politiques qui contrôlent l'accès au dataset à l'intérieur du data product. Par exemple, si le dataset contient des données personnelles identifiables (PII), une politique de confidentialité peut les rendre invisibles pour tous les utilisateurs ou bien uniquement pour certains.
- 8 **Contrats de données et SLO** : définissent les valeurs cibles pour l'objectif de service que le data product est supposé remplir. Par exemple, un SLO peut exiger que la qualité des données en termes de précision soit supérieure à 95 % ou que les données soient actualisées toutes les cinq minutes.
- 9 **Fonctions analytiques** : encapsulent une logique métier personnalisée supplémentaire pour rendre la fonction réutilisable. Un exemple de fonction sur un data product dans un service d'enregistrement des commandes pourrait être « commandes en attente », qui encapsulerait la logique complexe du domaine pour déterminer quelles commandes sont en attente de traitement. Les fonctions analytiques peuvent être exposées via un port de sortie exactement comme une API.
- 10 **Dépendances de l'infrastructure** : informations sur les exigences en matière d'infrastructure pour les datasets associés au data product. Par exemple, les datasets sont stockés sur une plateforme de données (comme Snowflake ou Databricks), avec le code associé (par exemple, la logique de transformation) nécessitant un environnement de calcul (par exemple, SQL) pour être exécuté.

Cycle de vie des data products

Les data products ne sont pas des constructions ponctuelles et figées. Comme c'est le cas pour des produits matériels, ils nécessitent une gestion et une surveillance continues. Chaque data product appartient à une équipe data spécialisée dans un domaine. Cette dernière le gère et est responsable de ses performances, c'est-à-dire qu'il doit générer de la valeur et voir son adoption croître. Elle se doit également de maintenir de son cycle de vie.



Les principales étapes du cycle de vie d'un data product sont les suivantes :

- 1 Conceptualisation** : ce processus commence par l'identification des besoins professionnels spécifiques au domaine, notamment la compréhension des utilisateurs, de leurs cas d'usage propres et de leur mode courant de consommation des données.
- 2 Conception** : l'étape suivante consiste à élaborer la solution, ce qui inclut l'identification des datasets nécessaires et de leurs sources, ainsi que la définition des caractéristiques du data product en termes de fonctionnalité, d'interactivité, d'approche de la présentation, etc.
- 3 Ingénierie** : l'étape de la conception est suivie de celle de la création de l'infrastructure technique et des fonctionnalités du data product. Il s'agit de rassembler et d'intégrer les datasets bruts depuis leurs sources, d'assurer la qualité des données et de les transformer.
- 4 Déploiement** : cette étape consiste à compléter la documentation du produit, à en désigner le propriétaire, à l'intégrer aux systèmes d'entreprise ou aux workflows de consommation existants pour un accès facile, et à activer le data product ou à le rendre prêt à être consommé.
- 5 Mise sur le marché** : une fois l'activation terminée, il est important de sensibiliser ceux qui vont y recourir et de promouvoir l'utilisation de la technologie. Cela implique de former les consommateurs et de mettre en place des canaux de communication pour recueillir leurs commentaires.
- 6 Utilisation** : si le lancement est réussi, les consommateurs commenceront à interagir avec le data product pour le rechercher, le trouver, le comprendre et y accéder. Cette étape se concentre sur le monitoring de l'utilisation et la collecte des commentaires. Il s'agit généralement de recueillir des données sur la manière dont les utilisateurs interagissent avec le data product, sur ses performances et sur sa capacité à répondre aux besoins de l'entreprise en fonction du domaine spécifié.
- 7 Maintenance** : les data products ne sont pas des entités statiques ; ils doivent être améliorés pour garantir une valeur continue. Les commentaires des utilisateurs recueillis et l'évolution des besoins de l'entreprise permettent de proposer de nouvelles dimensions de données, caractéristiques et fonctionnalités à ajouter dans la prochaine version du data product.

En suivant ces étapes du cycle de vie, les organisations peuvent créer des data products qui fournissent des insights pertinents, favorisent la prise de décisions data-driven et permettent d'atteindre leurs objectifs métier.

Ce que le data product N'EST PAS

L'univers des données est truffé de termes, dont l'usage n'est pas toujours cohérent. Exemple concret : l'expression « data product » est souvent confondue avec des termes comme « dataset » ou « catalogue de données », qui ne représentent pourtant pas la même chose. Cette section souligne certaines différences fondamentales entre ces concepts.



Bien qu'un dataset soit un élément fondamental de nombreux data products, il est essentiel de comprendre qu'il ne s'agit que d'un élément parmi d'autres. Un data product va bien au-delà du dataset brut.

LES DATASETS : LA MATIÈRE PREMIÈRE

Un dataset est en général un ensemble de points de données organisés, que ces dernières soient structurées ou non. En général, il s'agit d'informations brutes telles que des données sur les clients, des dossiers financiers ou des relevés de capteurs. On peut considérer les datasets comme de la matière première, tout comme l'est un tas de bois, qui a le potentiel de devenir précieuse, mais qui doit d'abord être travaillée. L'une des principales différences entre un dataset et un data product réside dans l'absence d'un contexte métier spécifique à un domaine.

DATA PRODUCTS : LES PRODUITS FINIS

Les data products sont des datasets méticuleusement conservés et traités, souvent associés aux éléments dont il a été question dans la section précédente, afin d'apporter une réelle valeur ajoutée.

Voici quelques différences entre les data products et les datasets :

- **Contexte et spécificité de domaine** : les data products sont adaptés pour répondre à des besoins métier ou des demandes utilisateurs spécifiques. Ils vont bien au-delà des données brutes et contiennent des informations contextuelles telles que les définitions, la traçabilité et les contrôles de qualité.
- **Facilité d'utilisation et accessibilité** : les data products sont conçus pour optimiser la valeur et faciliter la consommation. Ils sont traités et « conditionnés » de manière à être facilement utilisables par des consommateurs de données spécialisés dans un domaine, avec une expertise technique limitée.
- **Des insights exploitables** : les data products ne se contentent pas de présenter des données ; ils proposent des insights opérationnels, analytiques, prédictifs et prescriptifs sous forme de tableaux de bord de visualisation, de modèles analytiques préconstruits ou de liens d'intégration avec d'autres outils par le biais de SQL ou d'API.
- **Propriété et responsabilité** : les data products ont des liens de propriété sans équivoque, ce qui garantit la responsabilité des différents aspects de leur développement, de leur consommation et des commentaires faits à leur sujet au sein de l'organisation. La propriété des data products dépend du contexte métier, le propriétaire pouvant être le producteur qui génère le data product, le consommateur au sein de l'équipe du domaine responsable des cas d'usage adaptés ou un rôle nouvellement émergent connu sous le nom de « data product manager ».



Les data products transforment les datasets en assets précieux et consommables qui jouent un rôle important dans la prise de décisions avisées et la croissance de l'entreprise. La revue Harvard Business Review estime que « les entreprises qui traitent les données comme un produit peuvent réduire de 90 % le temps nécessaire à leur implémentation dans de nouveaux cas d'usage, diminuer de 30 % les coûts totaux de propriété et réduire les risques et le fardeau de la gouvernance des données ».⁴

⁴ « A Better Way to Put Your Data to Work », Harvard Business Review, juillet-août 2022.

CATALOGUE DE DATA PRODUCTS : LE HUB DES PRODUITS FINIS

Un catalogue de data products est un référentiel complet qui contient des data products et les métadonnées associées comme des descriptions, des tags, des versions, des étiquettes de propriété, etc. Ce catalogage aide à la découverte et à la compréhension des data products. Il s'adresse à la fois aux producteurs et aux consommateurs de données, en proposant des solutions sur mesure pour répondre à leurs besoins. Pour les data product managers, le catalogue fournit des outils pour configurer, mettre en œuvre et gérer le cycle de vie complet d'un data product, de sa création à sa fin de vie, en passant par son activation. Pour les consommateurs, le catalogue sert de marketplace, ce qui permet une expérience similaire à celle du shopping et favorise l'utilisation.

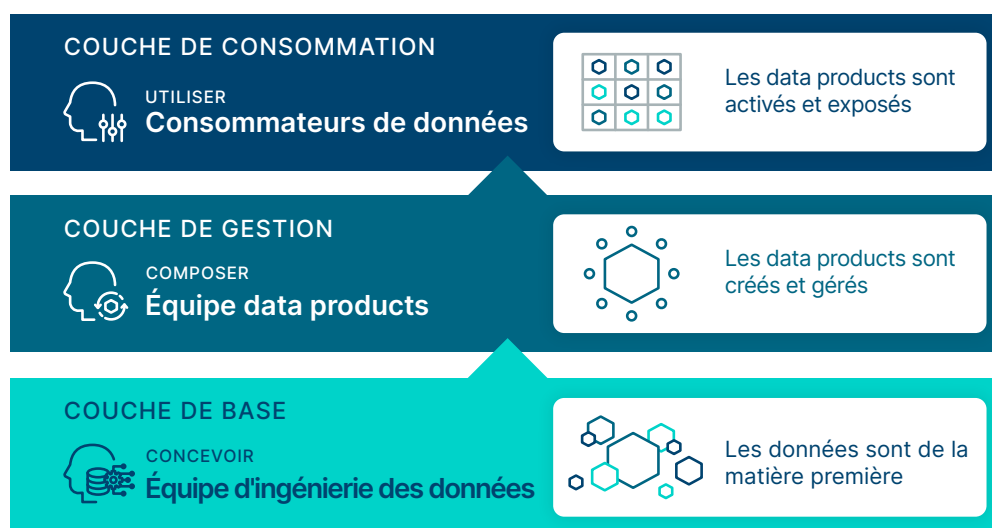
Le catalogue de data products se distingue du catalogue de données par les points suivants :

FONCTION	CATALOGUE DE DONNÉES	CATALOGUE DE DATA PRODUCTS
Mission	Renforcer la gouvernance des données	Accroître la valeur des données
Périmètre	Entreprise	Domaine spécifique
Modèle de gestion	Centralisé	Fédéré
Finalité	Visibilité, conformité, transparence	Facilité d'utilisation, confiance, capacité à être exploité
Contenu	Métadonnées de diverses natures, notamment techniques, opérationnelles et métier, relatives à chaque asset de donnée	Data products fiables, métadonnées du cycle de vie des data products, documentation des data products et propriété
Entités	De l'ordre de centaines de milliers	De l'ordre de centaines
Audience	Principalement experts en gouvernance des données	Producteurs et consommateurs de données

Le véritable enjeu dans le domaine de la gestion des données n'est pas seulement de créer des plateformes de données, mais aussi d'en assurer l'adoption effective. Les organisations investissent trop souvent des budgets considérables (parfois des millions de dollars) dans la création d'infrastructures de données, pour finalement se heurter à un problème de taille : **le partage des données entre producteurs et consommateurs**. Essayons de comprendre pourquoi : les résultats générés par de nombreuses plateformes technologiques manquent de fiabilité et ne sont pas adaptés aux objectifs métier spécifiques des consommateurs. **En fin de compte, vous avez besoin d'un catalogue de données et d'un catalogue de data products dans votre entreprise.** Ils sont comparables aux différents rayons d'un magasin d'alimentation, l'un contenant les matières premières telles que la farine, le beurre et le sucre, et l'autre les produits finis tels que les gâteaux et les pâtisseries.

Catalogue de data products : couches et acteurs

Le catalogue de data products est le point central pour organiser des data products au sein d'une organisation. Il est structuré en trois couches, chacune adaptée à un ensemble différent de parties prenantes. Qu'il s'agisse de data engineers, de data stewards, de data analysts ou de nouveaux rôles tels que les data product managers, chaque groupe interagit de manière unique avec ces couches, en fonction de ses objectifs et de ses besoins. Cette approche permet de cloisonner les problématiques propres à chaque groupe d'utilisateurs, d'utiliser un degré d'abstraction approprié à leurs exigences, afin de faciliter l'adoption et la consommation.



LA COUCHE DE BASE

La couche de base du catalogue de data products est essentielle. Elle établit l'infrastructure technique et fournit les capacités qui permettent la prise en charge des couches suivantes.

Les principales capacités de cette couche sont :

- **Ingénierie des données** : déplacement et traitement des données provenant de diverses sources, notamment de fichiers, cloud data warehouses, data lakes et streams de données en temps réel.
- **Gestion des métadonnées** : collecte et organisation d'informations importantes sur les données, telles que le lignage des données, les dictionnaires de données, les schémas, etc.
- **Sécurité et contrôle d'accès** : mise en place de mesures de sécurité pour protéger les données et s'assurer que seuls les utilisateurs autorisés peuvent y accéder.
- **Qualité des données et suivi du lignage** : monitoring de la qualité des données pour garantir l'exactitude et le suivi du lignage des données afin d'en assurer la fiabilité.

La couche de base sert principalement à l'équipe chargée de la plateforme de données, y compris les rôles techniques pratiques tels que les data engineers, les data stewards et les data architects. L'objectif est de développer les capacités de cette couche en libre-service afin que les utilisateurs non techniques puissent les solliciter facilement. Les capacités requises dans la couche de base peuvent être trouvées dans divers outils de gestion des métadonnées, d'ingestion des données, de transformation des données, de qualité des données et de sécurité des données.

LA COUCHE DE GESTION

La couche de gestion des data products possède les capacités nécessaires pour répondre aux besoins du rôle émergent du data product manager. Cette couche permet de développer, gérer et maintenir des data products fiables et spécifiques à un domaine. Elle s'appuie sur les services du niveau précédent pour inventorier, déplacer, transformer et sécuriser les assets de données. Elle sert de pont entre les métadonnées brutes résidant dans la couche de base et les insights exploitables que les consommateurs cherchent à obtenir à partir de leurs données.

Les principales capacités de cette couche sont :

- **Composition d'un data product** : capacités à concevoir un data product, à définir ses entrées et ses sorties, et à configurer tous ses différents éléments.
- **Déploiement d'un data product** : gestion des versions et du workflow pour activer un data product lorsqu'il est prêt à être consommé par les utilisateurs finaux.
- **Maintenance du data product** : suivi de l'utilisation, collecte des commentaires, modifications et mises à jour, et saisie de toutes les étapes du cycle de vie du data product décrit dans les sections précédentes.

La couche de gestion des data products sert principalement à l'équipe de gestion du produit, en particulier aux data product managers qui pilotent le cycle de vie des data products afin d'atteindre les objectifs de l'entreprise.

LA COUCHE DE CONSOMMATION (OU LA MARKETPLACE)

La couche de consommation se situe au sommet du catalogue de data products et représente la partie du système orientée vers le consommateur. Elle permet l'exploration et la consommation de données en libre-service, favorisant ainsi l'instauration d'une culture data-driven au sein de l'entreprise.

Les principales capacités de cette couche sont :

- **Recherche et découverte** : des interfaces intuitives permettent aux utilisateurs d'explorer et de trouver facilement les data products qui répondent le mieux à leurs besoins.
- **Apprentissage et confiance** : ici, les consommateurs de données ont accès à des informations leur permettant une meilleure compréhension du data product. Ils peuvent examiner des métadonnées et des échantillons de données, apprendre à l'utiliser en examinant des exemples de code et la documentation sur les données et renforcer leur confiance en passant en revue la qualité et l'historique des données.
- **Accès et autorisations** : cette couche contrôle qui peut accéder à des data products spécifiques, assurant ainsi la sécurité des données et la conformité avec les politiques de gouvernance des données.
- **Interfaces de consommation des données** : les API, les outils de visualisation ou même les points d'accès directs aux bases de données fournissent aux utilisateurs les moyens de consommer les data products. Selon le produit, différentes interfaces peuvent être disponibles pour répondre aux divers besoins et compétences des utilisateurs.

La couche de consommation des données sert principalement aux analystes métier, aux data scientists, aux utilisateurs opérationnels et à d'autres décideurs clés de l'entreprise.

AVANTAGES DES DATA PRODUCTS POUR LES DIFFÉRENTS ACTEURS DE L'ÉCOSYSTÈME DES DONNÉES

Les données sont les assets les plus précieux d'une organisation, mais leur véritable potentiel ne peut être exploité que si elles sont facilement accessibles, bien comprises et utilisées de manière efficace. Là est la finalité des data products : ils transforment les données brutes en assets consommables et fiables qui donnent aux parties prenantes de l'organisation les moyens d'agir.

Voyons le rôle que jouent les data products auprès des acteurs clés de l'écosystème des données :



Pour les **équipes d'ingénierie des données**, les data products sont un langage commun, un accord clairement défini, qui apporte une structure, et précise les relations entre les producteurs et les consommateurs de données. Cela permet aux producteurs de générer des data products réutilisables de manière itérative, réduisant ainsi le coût total de possession. En plus de cela, l'équipe d'ingénierie des données n'a plus besoin de concevoir de grands pipelines de données centralisés et complexes. Elle peut se concentrer sur le développement de pipelines plus petits et indépendants dédiés à l'implémentation de chaque produit.



Pour les **équipes en charge de la gouvernance des données**, les data products deviennent un point d'audit central et, comme ils sont fiables et gouvernés, il est plus simple de minimiser les risques. L'équipe en charge de la gouvernance définit et centralise les politiques et les normes que les différentes équipes data products appliquent ensuite de manière fédérée.



Pour les **data product managers**, les data products permettent de mieux aligner les livraisons des producteurs aux besoins des consommateurs de données, ce qui favorise la collaboration et encadre la responsabilité dans l'ensemble de l'écosystème des données. Cela renforce le sentiment de propriété et de responsabilité.



Pour les **consommateurs de données**, les data products sont des assets de données fiables, spécifiques à un domaine, qui fournissent exactement ce qui est nécessaire pour maximiser la valeur de l'entreprise. Cela permet de réduire les tâches redondantes et d'obtenir des résultats plus rapides.

Tous ces acteurs peuvent s'appuyer sur un catalogue de data products pour exploiter efficacement toute la valeur de leurs assets de données, favoriser des prises de décisions data-driven et atteindre leurs objectifs métier.

« Dans le secteur du crédit immobilier qui est le nôtre, nous considérons que les data products ne se limitent pas à des chiffres et à des algorithmes, mais qu'ils sont de véritables vecteurs de transformation. En créant des data products en libre-service, notre objectif est de donner les moyens à nos agents de proposer la solution la mieux adaptée à chaque emprunteur, tout en permettant aux équipes financières de mieux évaluer les coûts et de contrôler les risques. Pour nous, l'approche consistant à exploiter les data products dans nos activités est une véritable révolution métier. »

Julia Fryk

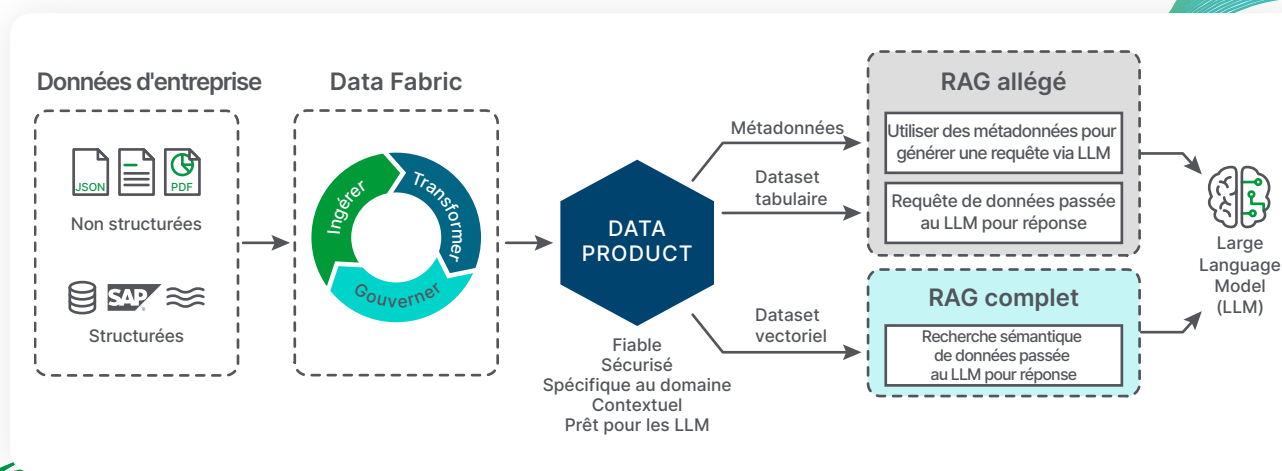
Principal Data Architect, Waterstone Mortgage

Fournir des données contextuelles aux LLM grâce à des data products

L'IA générative offre de formidables possibilités d'amélioration de la productivité dans tous les domaines de l'entreprise. Au cœur de cette technologie, se trouvent des modèles fondamentaux, des LLM tels que GPT-3, Anthropic, Llama et BERT, qui ont été pré-entraînés sur un corpus volumineux de données textuelles (pages web, documents, etc.). Cependant, les LLM de base ne comprennent pas les données internes, spécifiques à l'entreprise. C'est pourquoi ils doivent d'abord être nourris de données de l'organisation.

L'une des techniques les plus efficaces pour fournir des données d'entreprise aux LLM est un mécanisme appelé RAG (retrieval-augmented generation). Les données contextuelles liées au prompt d'un utilisateur (une question) sont **extraites**, puis ce même prompt est **enrichi** de ces données et transmis au LLM qui génère une réponse. Il existe deux méthodes RAG : la méthode allégée et la méthode complète.

- Dans le RAG allégé (pour les données structurées), l'application RAG transmet les métadonnées au LLM pour générer une requête SQL. La requête est ensuite exécutée sur le dataset tabulaire et les résultats sont transmis au LLM qui génère une réponse. Cette méthode est utilisée lorsqu'une réponse explicite est requise à partir des données fournies.
- Dans le cas d'un RAG complet (pour des données structurées ou non structurées), les données sont d'abord vectorisées. L'application RAG effectue ensuite une recherche sémantique sur ces données vectorisées pour les extraire. Elles sont ensuite transmises au LLM qui génère une réponse.



Dans les deux cas, il est important de disposer de données contextuelles d'excellente qualité, nettoyées pour le RAG. **Les data products fournissent des données contextuelles et spécifiques à un domaine de différentes manières pour la consommation.** Ils contiennent des métadonnées enrichies (métier, techniques et spécifiques à un domaine) et des datasets soigneusement traités (fichiers, tableaux ou vectorisés) pour la consommation. Ils se trouvent donc être le meilleur moyen de mettre en œuvre des applications basées sur les RAG.

La création d'une application RAG commence par une recherche dans le catalogue/la marketplace de data products pour trouver celui à utiliser en fonction de la question posée. Par exemple, une question relative aux ventes trouvera la meilleure réponse à l'aide d'un data product « Opportunité » provenant du CRM, sur la base des métadonnées détaillées qui lui sont associées. Par ailleurs, une question relative à une commande s'appuiera sur un data product « Enregistrement de commandes » provenant de SAP, sur la base des métadonnées détaillées qui lui sont associées. Une fois le data product identifié, l'application RAG peut directement interroger son dataset, qu'il soit tabulaire ou vectoriel, en fonction du cas d'usage.

Conclusion

Alors que les données sont résolument le moteur des entreprises modernes, les data products deviennent incontournables **pour générer toute la valeur des données** et accélérer les résultats des entreprises, tout en **améliorant la confiance et la gouvernance des données**. Qlik s'engage à vous accompagner dans cette démarche, à mesure que vous bâtissez un socle de données dédiées à des capacités en libre-service. Il s'agit de permettre à vos équipes de créer et de gérer des data products, tout en donnant aux consommateurs de données la possibilité de les rechercher, de les découvrir et de les utiliser rapidement pour remplir les objectifs de votre organisation.

Une dernière remarque, mais non des moindres : si les data products sont des outils puissants, ils ne constituent pas une solution miracle. La véritable puissance des data products réside dans l'identification des besoins de l'entreprise, dans une stratégie de données fédérée bien définie, dans la garantie d'une gouvernance efficace et dans la collaboration entre les équipes.

Envie de révéler toute la valeur de vos données ?

En savoir plus

« Pour les organisations, créer des data products répondant aux besoins de domaines particuliers, au-delà des questions d'intégration et de qualité des données, constitue aujourd'hui un enjeu majeur. Les offres de Qlik en matière de qualité et de gouvernance des données incarnent le concept de data products et témoignent d'une maturité qui correspond aux exigences des clients d'aujourd'hui. »

Stewart Bond

VP, Data Intelligence and Integration Software, IDC



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