LIVRE BLANC

Smartdata
by Ntico

Construction d'une Self Serve Data Platform

I. Introduction

Qu'est-ce que le Data Mesh ?

Le Data Mesh est une architecture et une méthodologie qui prônent le traitement des données comme **un produit à part entière**.

Un des axes majeurs dans la mise en place de cette stratégie consiste à **décentraliser la gestion des données** tout en maintenant des standards élevés de gouvernance et de qualité.

Le Data Mesh repose sur quatre principes fondamentaux :

- Le Domain Ownership
- La notion de Data Product
- Une gouvernance fédérée
- **Une Self Serve Data Platform**

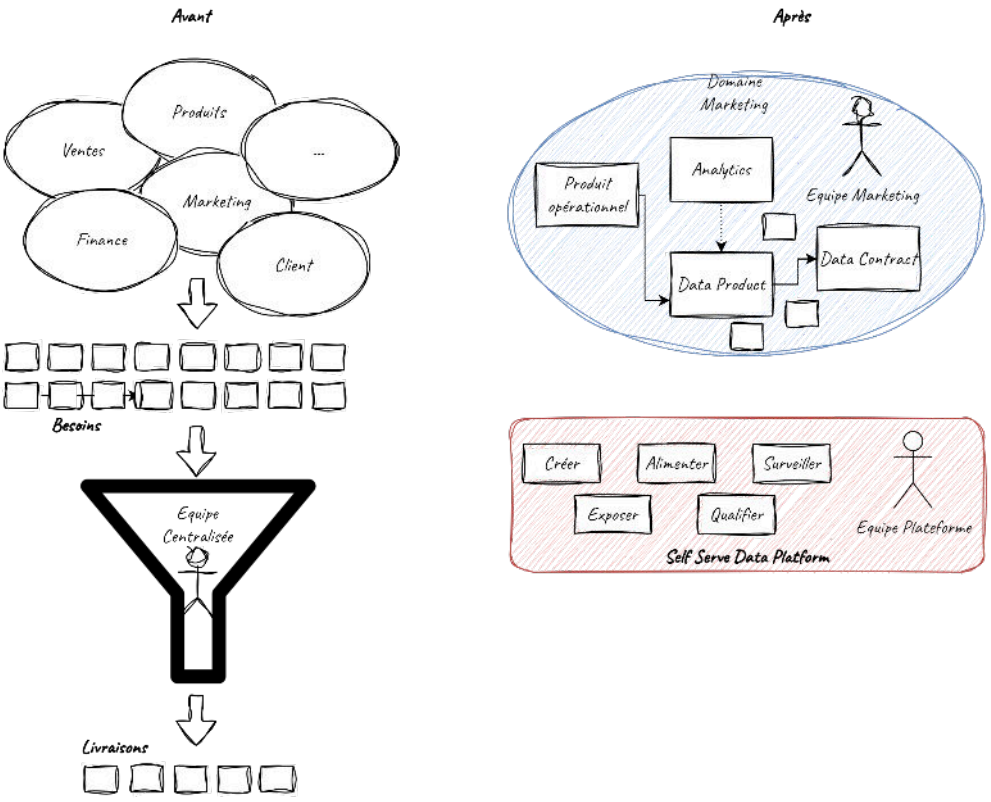
Ce document vise à présenter les points clés de la construction d'une telle plateforme, afin de vous donner quelques pistes pour soutenir vos ambitions autour de l'implémentation de Data Mesh au sein de votre organisation.

Pourquoi mettre en place une Self Serve Data Platform ?

Dans un environnement en **constante évolution**, les entreprises doivent s'adapter rapidement aux nouvelles demandes du marché et les équipes data sont alors sur-sollicitées.

La centralisation de ces ressources, un temps fortement favorisée afin de créer de vrais **centres de compétences** performants et revêtant d'un **niveau d'expertise** permettant de gérer n'importe quelle complexité projet, crée désormais parfois des goulets d'étranglement et de l'inefficacité, augmentant le Time To Delivery des projets.

Une plateforme de données en libre-service permet aux différentes équipes dans les domaines métier et IT de **gérer leurs propres ensembles de données**, ce qui favorise la **décentralisation et l'autonomie** tout en maintenant une gouvernance centrale efficace et un Time To Market mesuré.



Lorsqu'une stratégie est mise en place sans coordination, chaque équipe adopte ses propres outils et méthodes. Cela entraîne une multiplication des approches **avec autant de modes de gestion spécifiques qu'il y a d'intervenants**, favorisant ainsi un déficit d'interopérabilité flagrant combiné à un chaos général lorsqu'il s'agit de comprendre le système dans sa globalité.

L'objectif est donc de **limiter les compétences** à acquérir pour les différentes parties prenantes, mais aussi de réduire les coûts d'acquisition, de formation, de mise en œuvre et de maintenance. En standardisant les outils et les méthodes, les entreprises peuvent optimiser leurs ressources, diminuer les coûts opérationnels et garantir **une meilleure intégration des systèmes**. Cette harmonisation facilite également la montée en compétence des équipes, qui peuvent se concentrer sur **des outils et des processus communs**, réduisant ainsi le besoin de nombreuses formations spécifiques et coûteuses.

Enfin, il sera d'autant plus facile de faire évoluer cette plateforme si elle est centralisée, permettant ainsi une gestion cohérente et une adaptation rapide aux nouvelles exigences du marché. La centralisation offre une vision globale et intégrée des données, facilitant l'implémentation de nouvelles technologies, l'amélioration continue des processus et l'alignement stratégique avec les objectifs de l'entreprise. Une plateforme centralisée et bien gouvernée constitue ainsi le **socle d'une organisation agile**, capable de **s'adapter** en permanence et de **tirer parti des opportunités offertes** par l'évolution rapide des technologies et des marchés.

Présentation de la démarche

Dans un premier temps, nous détaillerons **l'architecture globale** d'une Self Serve DataPlatform, ainsi que la nécessité de penser Plateforme, au service des utilisateurs.

Ensuite, nous découvrirons de quelle façon il est possible de soutenir les équipes qui produisent les Data Products dans le management de leurs produits et de leurs pipelines de données.

En troisième lieu, nous allons décrire les mécanismes qui permettront aux consommateurs de ces produits d'explorer le patrimoine de données et de le valoriser.

De façon générale, il est important de noter qu'il est rigoureusement recommandé de **prioriser la mise en place de ces concepts selon vos besoins**, en prenant en considération les attentes de vos utilisateurs, en commençant par les équipes les plus avancées et à même de comprendre et mettre en oeuvre rapidement les outils que vous mettraient à disposition dans ce contexte. Cela permettra en outre de valider à chaque étape les décisions d'implémentation qui auront été prises

II. Architecture globale de la Self Serve Data Platform

Définition des attentes

Il est bien évidemment essentiel, avant de commencer quoique ce soit, de se questionner sur **les objectifs à atteindre**, et donc de définir les attentes de ses utilisateurs.
Les questions que l'on se doit se poser peuvent paraître évidentes, mais les attentes des uns ne sont pas celles des autres.

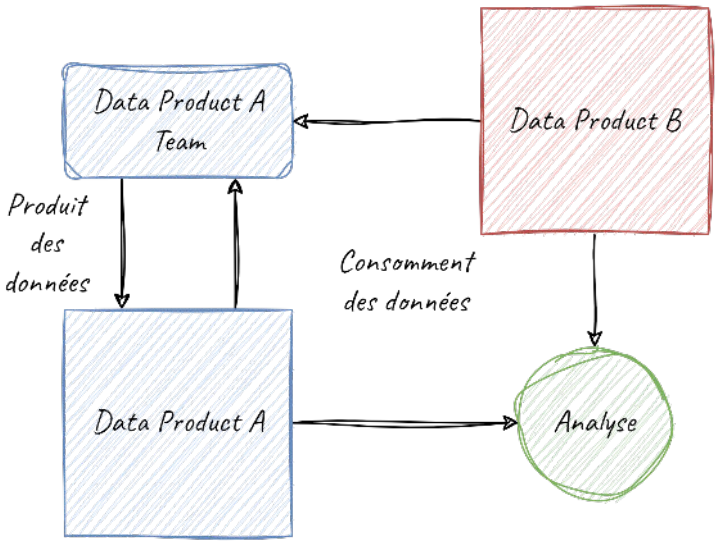
Nous devons alors bien **identifier les différentes parties prenantes** impliquées dans ce que sera la future gestion des données. **Déterminer** avec elles les différentes **fonctionnalités** qui sont attendues et les **interactions** qu'il faudra être en mesure de gérer grâce à la SSDP.

En outre, il conviendra également de spécifier le niveau d'automatisation des gestes les plus courants ou les plus risqués afin d'obtenir le plus de bénéfices possibles de cette plateformisation, que ce soit en matière de gouvernance ou d'ingénierie.

Dans la plupart des cas, les personas qui devront être engagés dans la démarche seront principalement :

- Les Data Owners ou Data Products Owners
- Les Data Engineers et apparentés (puisque les Software engineers seront sans doute mis à contribution dans le nouveau modèle d'organisation.
- Les Analysts (Data ou Business)

Ils pourront alors tous être à la fois, dans certains cas, Producteurs de données, et dans d'autres cas, Consommateurs de Données.
Dans tous les cas, ils interagiront avec des Data Products, au travers de la SSDP que vous devez mettre en place.

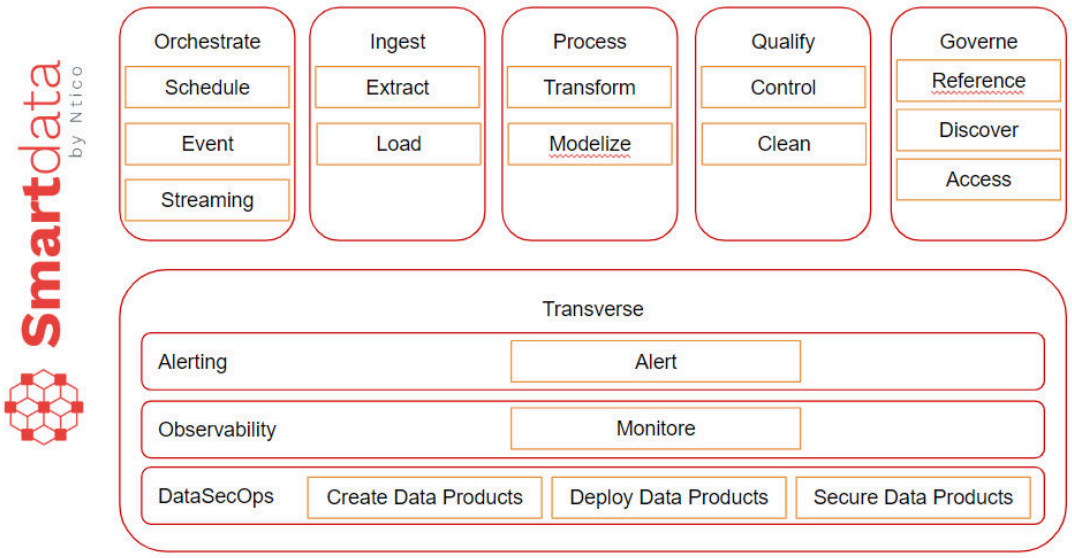


SSDP Architecture Building Blocks

Le périmètre est vaste et le champ fonctionnel très étiré. En effet, les besoins fonctionnels à couvrir, déterminés je le rappelle en concertation avec vos utilisateurs, concernent potentiellement **tout le pipeline de données**.

Tout cela nous conduit à penser **l'architecture globale de la SSDP**. Et il nous est apparu intéressant d'utiliser le framework d'architecture Togaf pour délimiter les contours de cette architecture en fonction de la couverture fonctionnelle à adresser.

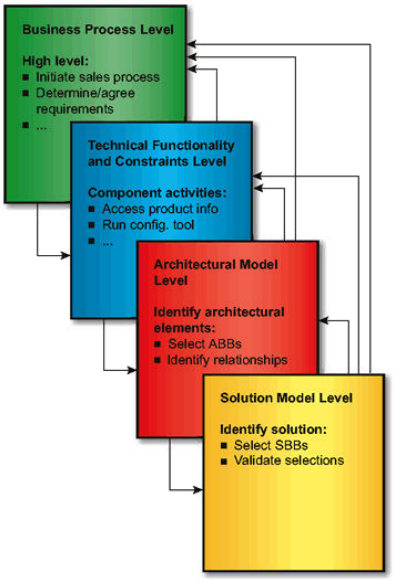
- **Architecture Building Blocks**



Au fur et à mesure que les utilisateurs seront onboardés sur la plateforme, **les besoins en ressources augmenteront**, les données demanderont toujours plus d'espaces de stockage et de capacités de traitement. La SSDP devra être pensée de sorte qu'elle puisse être **évolutive**, en termes de fonctionnalités et donc d'apport de valeur mais aussi de scalabilité.

Inévitablement, la conception de votre plateforme implique **des intéractions et des intégrations de plus en plus complexes**, comprendre le rôle de chaque composant devient essentiel.

Togaf nous aide grâce aux principes **ABB/SBB** (Architecture Building Blocks / Solution Building Blocks), en adoptant une approche centrée sur les **processus métier**.



Cela permet de définir des composants qui couvrent un périmètre d'activité, formant un modèle d'architecture menant à une **solution unique et cohérente**.

Cette approche garantit la **stabilité**, la **qualité** du service et une intégration optimale, offrant ainsi des **expériences utilisateur fluides**.

Ref : https://pubs.opengroup.org/architecture/togaf

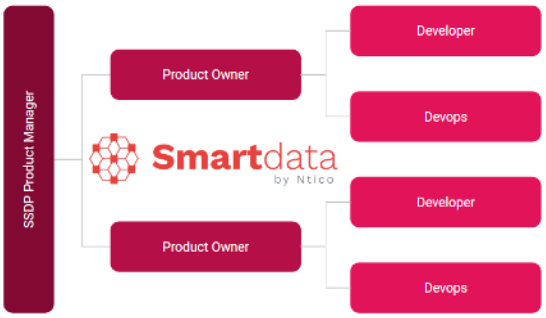
Organisation

Sur le plan humain, il est essentiel de constituer une équipe Plateforme solide.

Product Manager :
Nous vous conseillons d'inclure au minimum un Product Manager **expérimenté** dans ce type de produits, que ce soit dans la data ou dans des secteurs comme le e-commerce, où l'expérience des plateformes est acquise.

Product Owners :
Il vous faut également des Product Owners pour piloter la **roadmap** et le développement des produits qui composeront la SSDP.

Équipe de développement :
Enfin, des équipes de développement dédiées seront nécessaires pour les produits nécessitant un développement en **interne**.



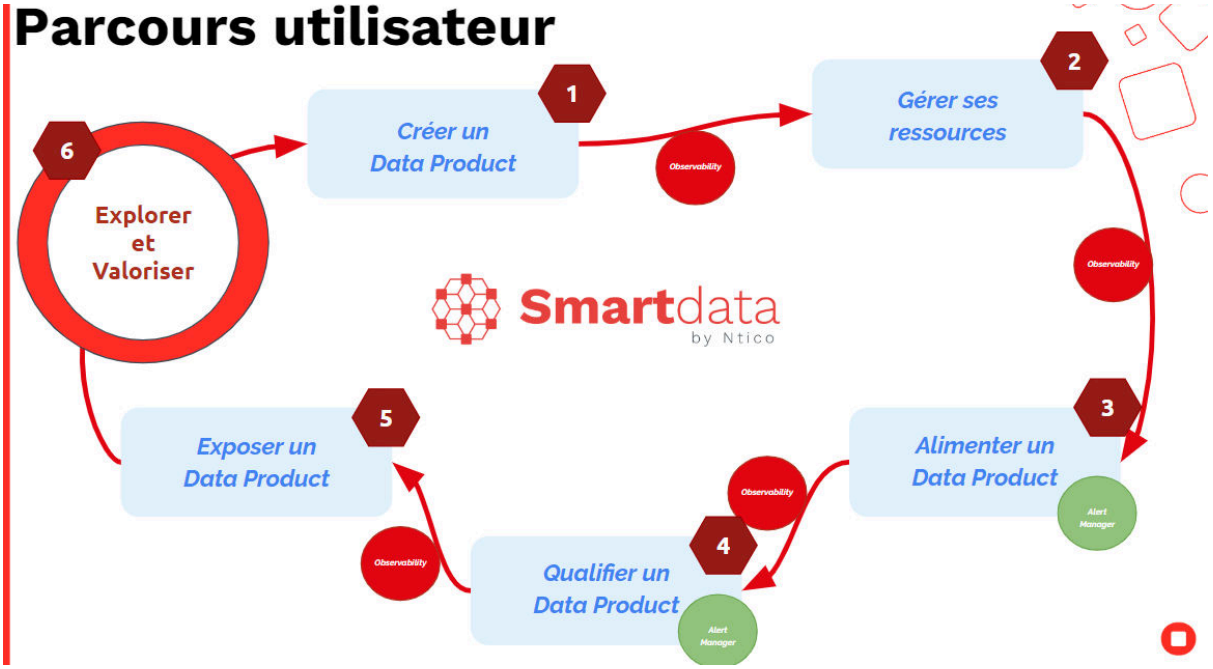
Sur ce dernier point, les choix d'outils dépendront de la complexité de vos besoins, des solutions disponibles qui couvrent tout ou une partie de ces besoins, ainsi que des contraintes organisationnelles, budgétaires et stratégiques liées aux processus métier à adresser.

Dans la suite de ce document, nous reviendrons sur ces aspects lorsque nous jugeons nécessaire de se pencher sur la question.

III. Gestion des Data Products

Vous vous apercevez désormais que plusieurs produits socle devront constituer cette plateforme, adressant chacun des fonctionnalités différentes qui serviront à garantir le bon déroulement des opérations liées aux Data Products, le tout devant s'interconnecter afin de fonctionner en mode plateforme, faisant ainsi bénéficier vos "clients" d'une expérience utilisateur sans couture, et rendant ainsi les interactions entre les différentes parties prenantes plus fluides.

Nous avons choisi, dans le cadre de la rédaction de ce document, de présenter l'implémentation de ce type de plateforme sous la forme du parcours d'un utilisateur sur le chemin de la conception d'un Data Product, de bout en bout.



Création d'un Data Product

Il y a bien un début à tout.

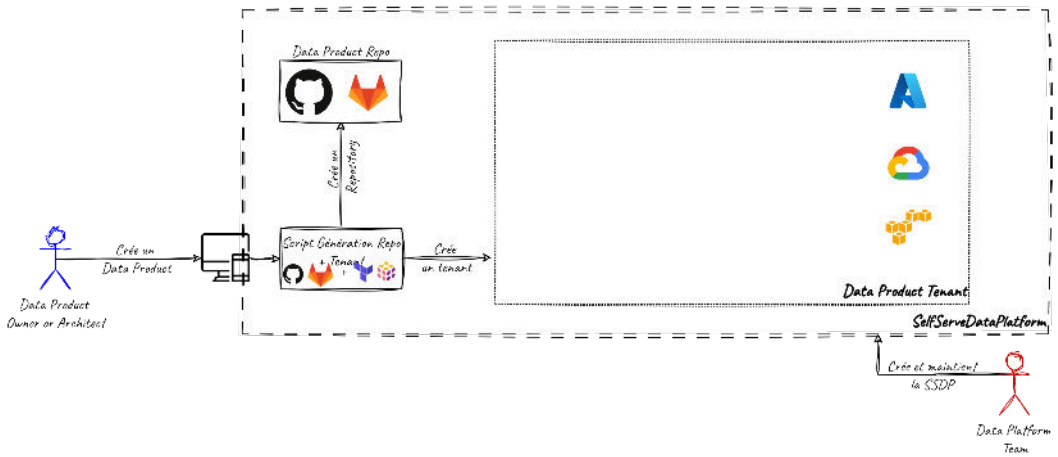
Et la première étape consiste à créer un Data Product, qu'il faudra ensuite être en mesure de gérer, selon les standards dictés par les politiques dressées par la Gouvernance Fédérée de l'entreprise. Le respect de ces politiques, énumérées ci-dessous, que l'on pourrait aborder dans un autre document tant il y a à dire, garantit que chaque équipe soit en mesure de créer, découvrir, comprendre et utiliser chacun des Data Products ainsi mis à disposition.

Grâce à un outillage dédié à la création de Data Products, il est possible de faciliter voire automatiser une partie de la mise en œuvre de ces politiques.

Cette étape est cruciale dans le déploiement de la plateforme, car elle va dessiner tout le processus de gestion des Data Products et garantir un cadre de travail clair et efficace.

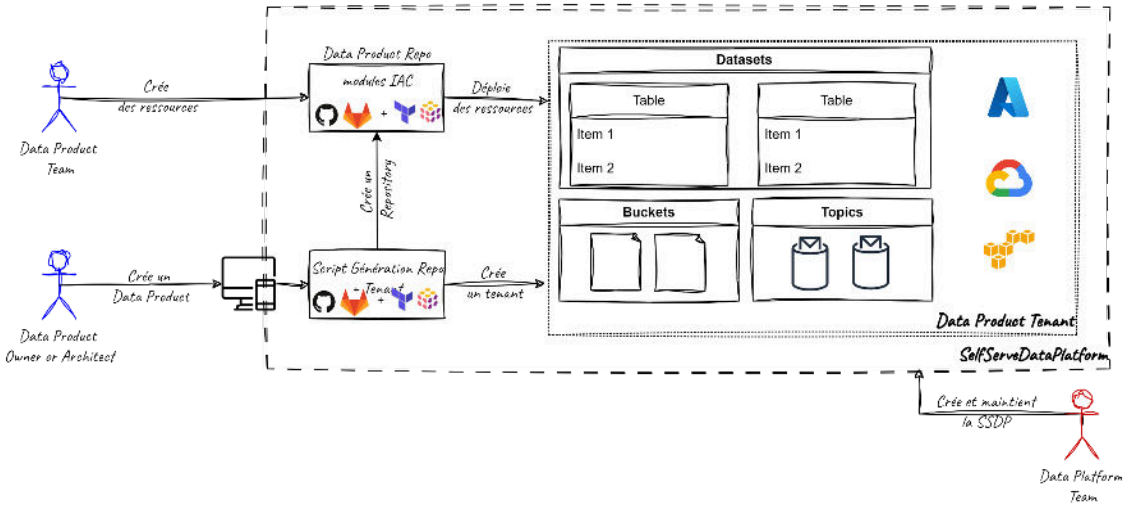
Afin de permettre de créer des Data Products standardisés et résilients, il convient de proposer aux utilisateurs une interface qui va automatiser sinon faciliter les gestes de création des environnements techniques.

L'idéal est de générer, grâce aux informations saisies sur un simple formulaire en libre service, et sur base de modèles de repositories et de tenants préconfigurés sur votre provider Cloud préféré, des landing zones où les utilisateurs peuvent héberger leurs Data Products. Cela permet d'améliorer significativement le Time to Delivery, en fournissant également des workflows de déploiement continu préconfigurés. Grâce à cet outil, les développeurs n'auront plus à penser à la technique de la technique, mais pourront se concentrer sur la valeur fonctionnelle de leurs Data Products.

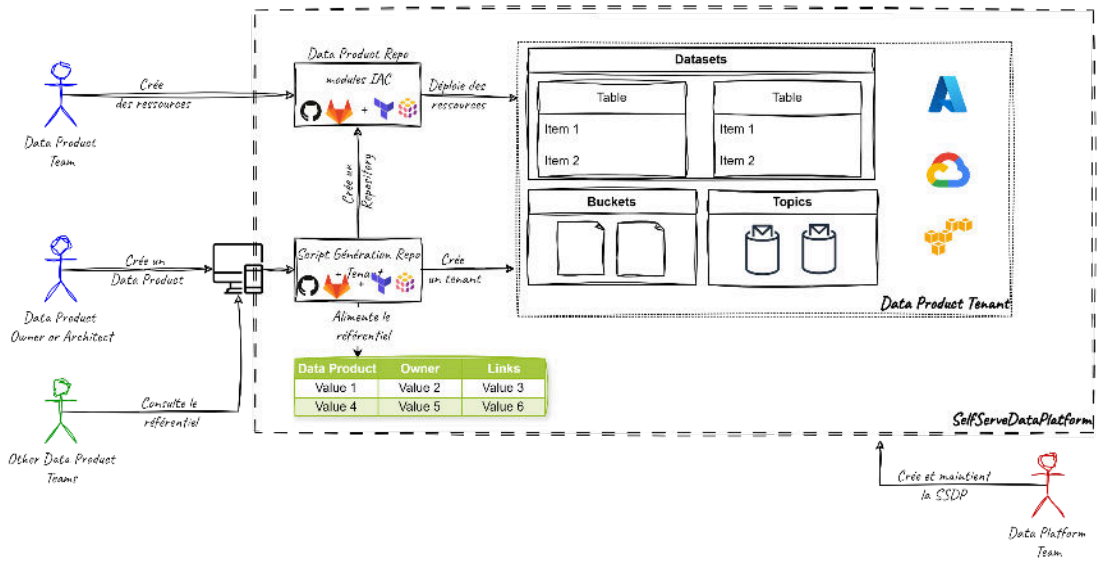


Gestion des ressources Data

Adossé à ce produit, il est important de proposer également des modules techniques de déploiement de ressource, de type Infrastructure As Code, que vous pouvez mettre à disposition de vos utilisateurs directement depuis le modèle de repository. Ces modules templatisés offriront des éléments supplémentaires à votre cadre de travail et ajouteront des moyens de standardisation de vos infrastructures et d'automatisation de vos politiques de gouvernance.
Des produits et services tels que Github / Gitlab ainsi que Terraform / Pulumi pourront vous être utiles dans la mise en oeuvre de ces modules.



En procédant de la sorte, ce produit peut également agir comme un référentiel, permettant de retrouver tous les repositories des Data Products et leur documentation. Il sera ainsi possible de développer l'inner sourcing au sein de l'ensemble des équipes Domaines, ce qui favorise l'amélioration de l'efficacité, le gain de temps, et donc l'amélioration des performances opérationnelles.



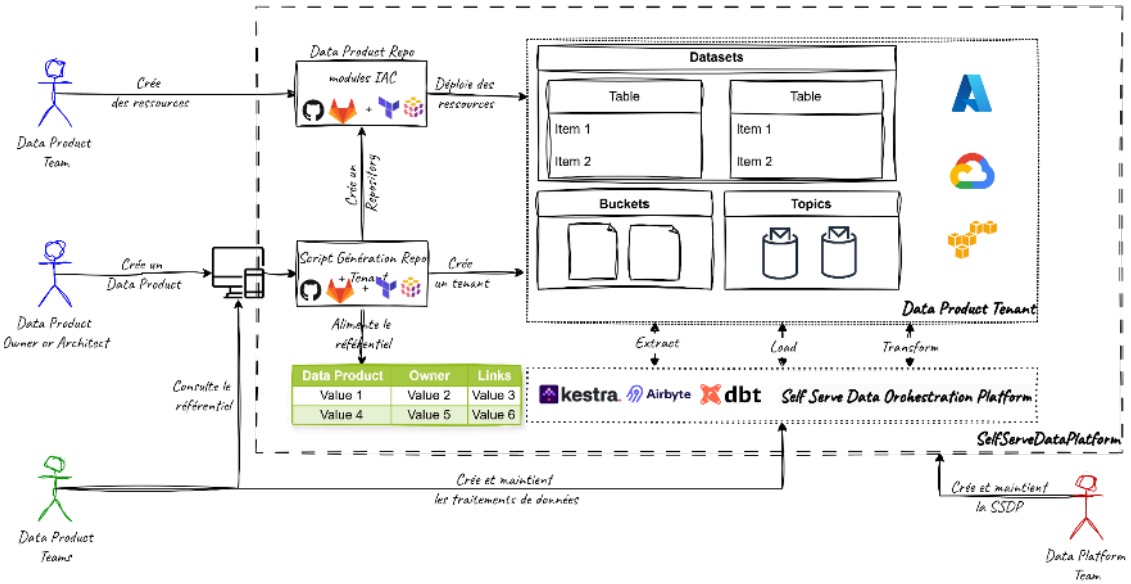
Alimentation d'un Data Product

Dès lors que votre Data Product est créé et que ses infrastructures sont disponibles, il peut être alimenté avec les données qui en seront sa principale composante.

Il existe généralement beaucoup de sources de données différentes auxquelles votre data product devra sûrement être connecté. Depuis des bases de données classiques, des systèmes de fichiers, en passant par les bases de données noSql et autres files de messages en tous genres, il sera important de choisir des outils qui disposent des connecteurs et fonctionnalités indispensables à la majorité de vos usages.

La plateforme de gestion de flux de données Kestra mérite alors toute votre attention, puisqu'en plus de disposer de plus de 540 connecteurs natifs et de la facilité qu'elle apporte dans la gestion de workflows pouvant adresser les besoins métiers les plus complexes, elle permet de traiter vos données aussi bien de façon planifiée, sur évènement mais aussi en temps réel, le tout pouvant être déployé en usant des meilleures pratiques d'industrialisation existantes.

Tous vos processus ELT deviennent alors extrêmement simples à mettre en place, pour toutes vos équipes Domaines, avec un minimum de compétences à acquérir puisque l'outil est agnostique des langages de programmation et des technologies sous jacentes grâce à son interface utilisateur simplifiante et sa syntaxe déclarative sous YAML.



Si votre organisation nécessite de pouvoir gérer des droits d'accès sur les pipelines de données, la version Entreprise Edition de Kestra permet de mettre en place des RBAC (Role Based Access Control) de façon très granulaire afin de respecter les contraintes d'ordre réglementaire qui peuvent parfois s'imposer dès lors qu'il est question de traiter de données client ou financières.

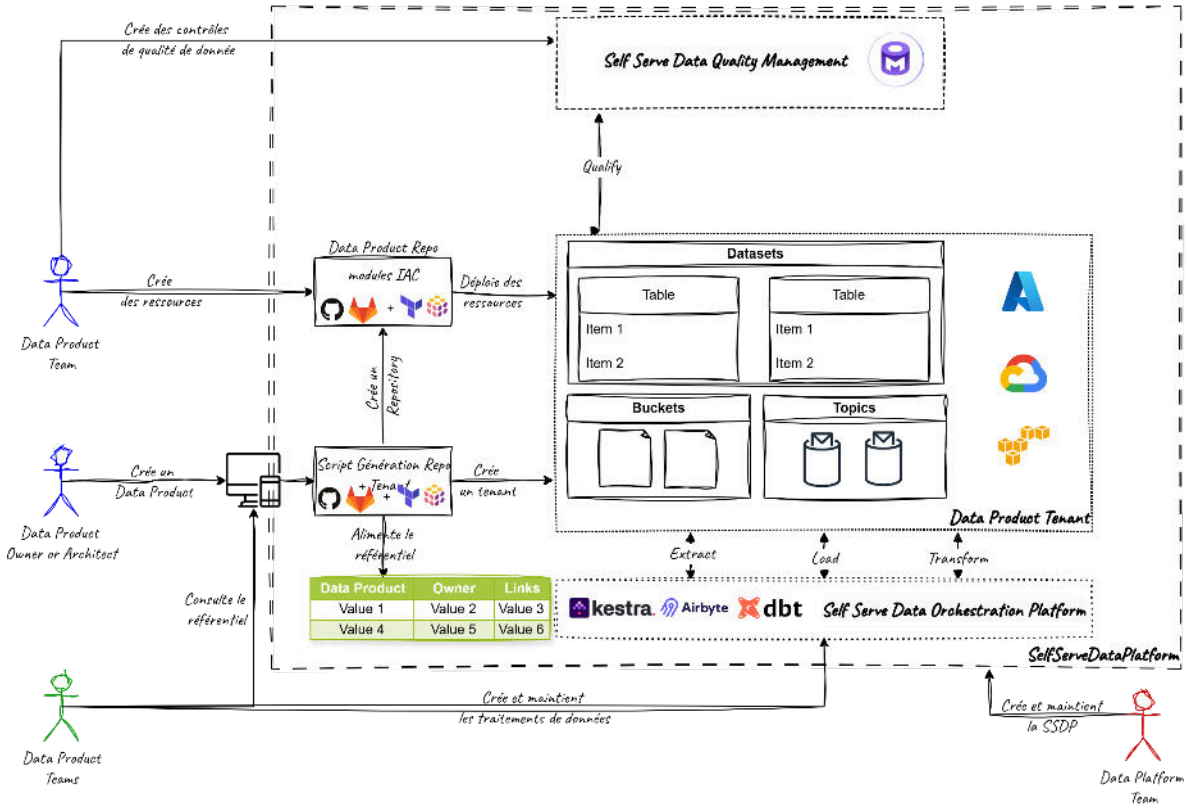
Combinée à des produits comme Airbyte ou Dbt, la solution Kestra gagne encore en puissance pour la gestion de tous vos pipelines !

Qualification des données d'un Data Product

Votre Data Product est désormais correctement alimenté, les données qui le constituent sont disponibles à l'endroit de leur stockage dans votre système de base de données.
Cependant, avant de mettre à disposition de vos consommateurs les données ainsi collectées, transformées voire générées, il est nécessaire de procéder au contrôle de la qualité de ces dernières.

En effet, nous l'avons dit précédemment, les données permettent de générer de la valeur. Cependant, une entreprise qui souhaite prendre des décisions grâce à la donnée pour piloter son business se doit de le faire avec des données fiables !
Il est important de comprendre que des décisions prises sur la base de données erronées peuvent nuire à votre activité. Le contrôle excessif des données par les équipes utilisatrices gaspille du temps et des ressources. De plus, la méfiance envers les données, une fois installée, est difficile à dissiper.

C'est pourquoi il est impératif de mettre en place un système de DataQualityManagement by Design. Contrôlez la qualité technique des données dès leur entrée dans la SSDP, voire même avant qu'elles ne soient entrées lorsque c'est possible. Ensuite, dès lors que des traitements sont opérés sur les données, procédez de nouveau à des contrôles qualité, cette fois plutôt sur le plan fonctionnel.



Le premier niveau de contrôle, technique, consiste à vérifier **la conformité des données** au regard du contrat d'interface qui vous engage vous et votre émetteur de données. Ces vérifications peuvent facilement être **automatisées** et rendues **obligatoires** pour chaque pipeline de collecte. Elles assureront vos équipes, vos opérations, et vos consommateurs que vos traitements se basent sur des données fiables, qui respectent les normes édictées pour la typologie de données échangées.

Le second niveau de contrôle, fonctionnel, permet à la team Domaine qui gère l'asset en question, de s'assurer que les traitements opérés sur la donnée collectée ne l'ont pas dénaturée et ainsi d'assurer à ses consommateurs que le niveau de qualité des données qu'ils exposent est optimal.

Il est donc primordial que les indicateurs de qualité de données soient accessibles en même temps que les données elles-mêmes. En tant que propriétaire d'un Data Product, vous devez donc exposer vos données, les contrôles effectués sur ces données ainsi que les résultats de ces contrôles.

Choisissez un outil qui permettra cette mise en avant des indicateurs auprès des consommateurs mais aussi qui soit assez simple à appréhender dans la mise en oeuvre des contrôles.
Pour cela, votre regard peut se porter sur la solution OpenMetadata.

L'outil propose des fonctionnalités spécifiques pour le suivi et la mesure de la qualité des données, y compris des indicateurs clé de performance (KPI), des règles de validation, et des rapports sur l'état de la qualité des données. Cela permet aux organisations de maintenir des standards élevés de qualité des données et de prendre des mesures correctives si nécessaire, cela même en temps réel grâce au système d'alerting et de monitoring intégré, permettant ainsi d'affecter le moins possible les processus métiers dépendants.

De plus, les utilisateurs peuvent **commenter, discuter et contribuer** à l'amélioration des données, favorisant ainsi un environnement où la qualité des données est une responsabilité partagée.

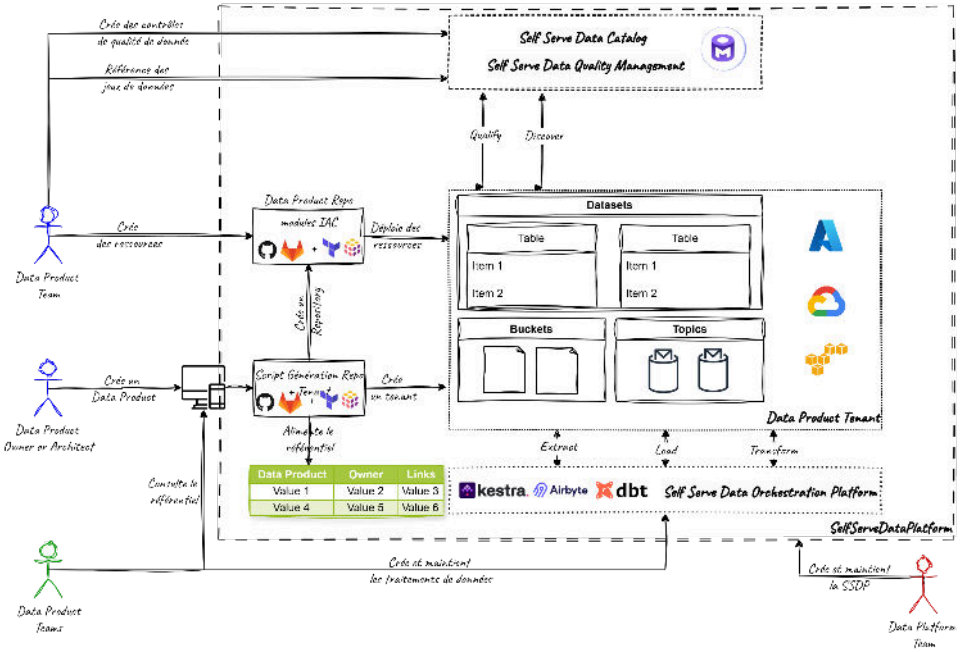
Exposition d'un Data Product

Tout est prêt ! Enfin presque !
Il va être question maintenant de donner une capacité à exposer les Data Products. Cette fonction est essentielle dans la SSDP puisque c'est elle qui rendra possible la consommation des assets auprès des utilisateurs finaux.

Au-delà d'être en mesure d'exposer les assets, il faudra surtout rendre possible leur documentation au travers d'un processus de référencement clair et guidé.
L'outil que vous allez choisir devra vous permettre de vous connecter facilement à vos différentes sources de données (Bases de données, files de messages, dashboards, ..) et ainsi recueillir nativement les métadonnées disponibles dans ces systèmes.
Par ailleurs, il sera important de laisser la possibilité à vos utilisateurs d'enrichir ces métadonnées avec des éléments spécifiques à leurs besoins, tout en restant dans un cadre de gouvernance maîtrisée.

Enfin, l'enjeu crucial qu'il faudra adresser dans l'exposition des data products sera sans nul doute l'accessibilité aux données et la gestion des droits afférente. Seuls les utilisateurs autorisés doivent pouvoir accéder aux données.

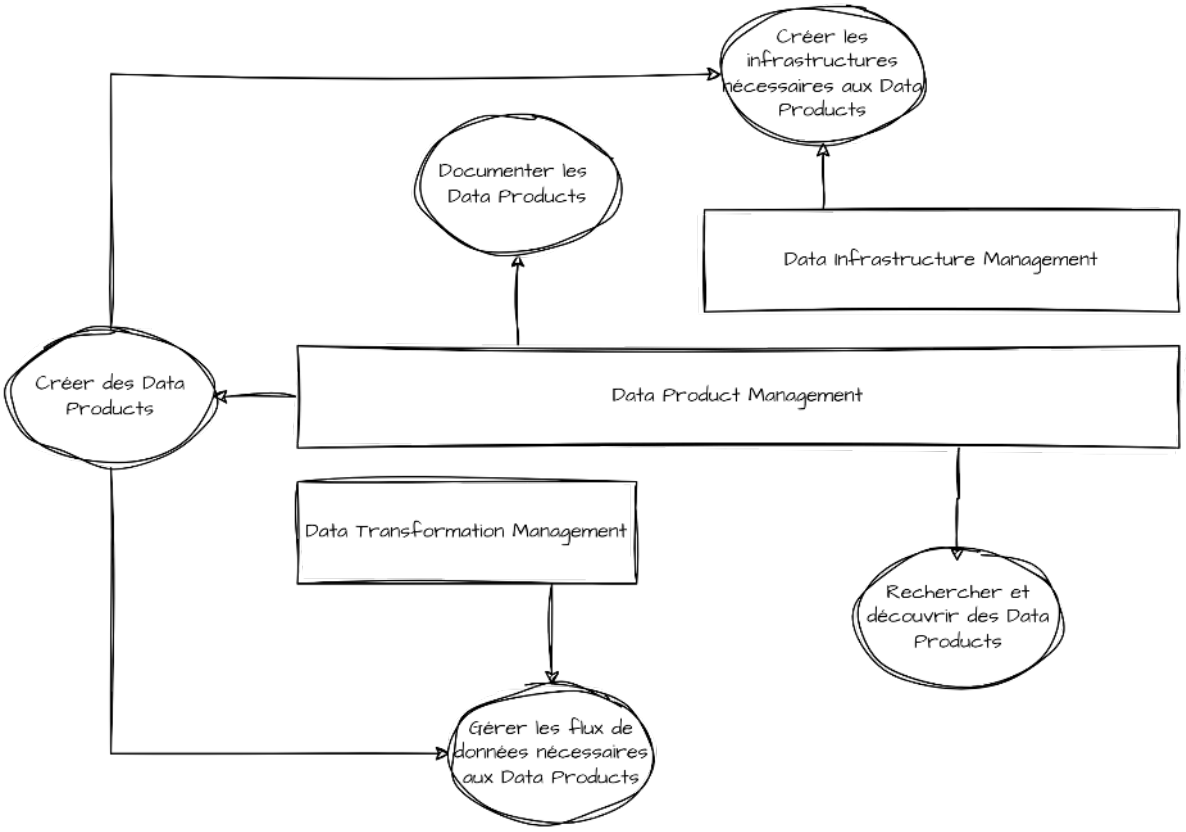
Là encore, OpenMetadata peut être un facilitateur dans la mise en place de ces fonctionnalités sur votre SSDP.
En effet, cette solution open source dispose d'atouts majeurs qui vous feront économiser en termes de budget et de ressources.



Grâce à ses connecteurs standards, vous pourrez découvrir rapidement tous les assets disponibles sur votre SSDP tels que les bases de données et les tables qui les composent, les topics, les pipelines d'un certain nombre d'outils du marché, les dashboards, les modèles de machine learning, que vos équipes domaines seront ensuite en mesure de documenter, référencer et exposer.
Il vous sera également possible de configurer des rôles et des permissions, liés à des teams et des domaines, qui permettront alors de gérer les accès aux différents data products.

IV. Exploration et Valorisation des Données

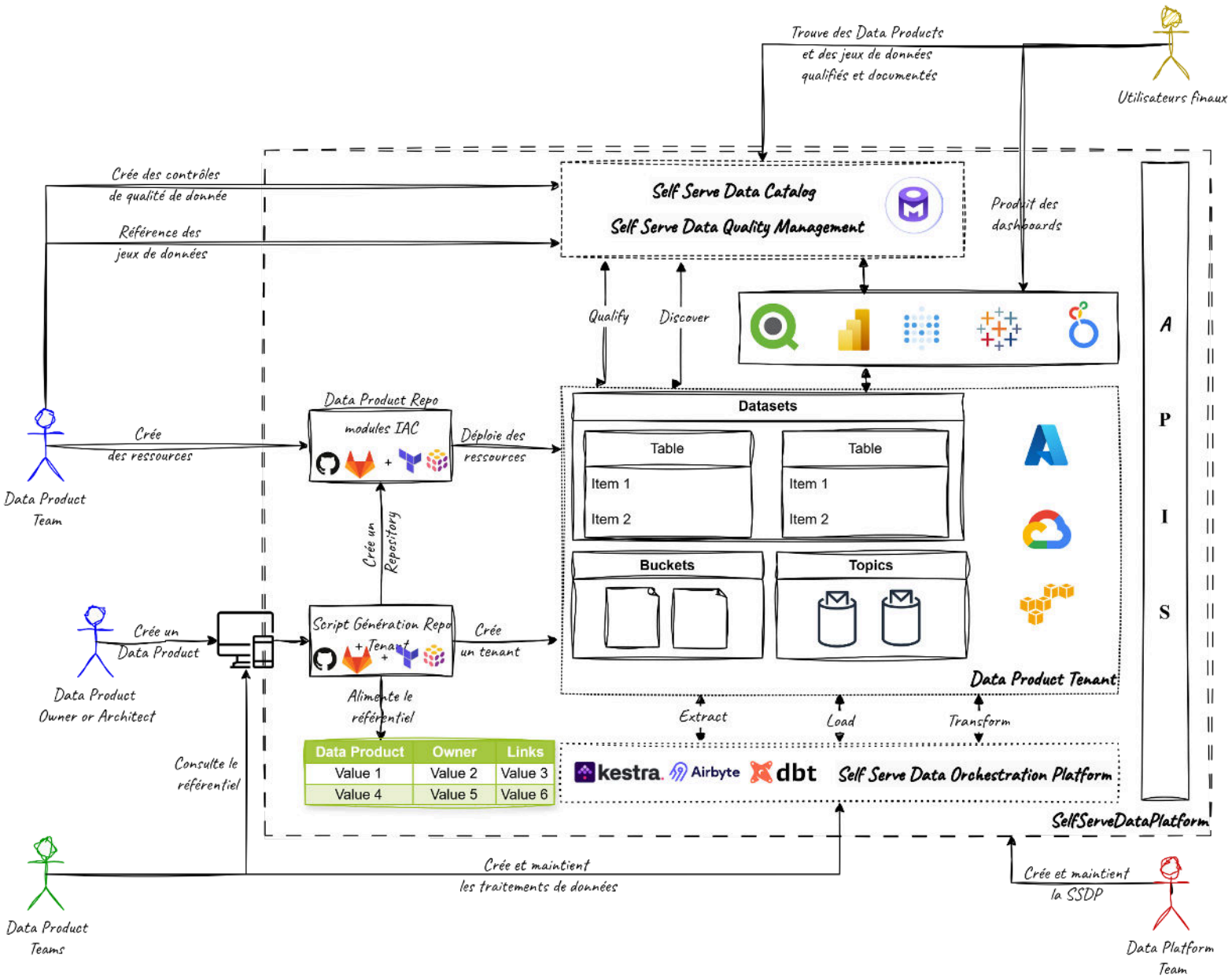
Vous avez tout au long de ce parcours construit votre SSDP sur un socle solide pour permettre à vos teams Domaines de créer des data products résilients, alimentés selon les besoins de vos métiers, rendant leurs données accessibles, découvrables et sécurisées.



Il est temps désormais pour tous les consommateurs de les exploiter et de les valoriser.
Grâce au Data Catalog mis à disposition, les différents utilisateurs de la plateforme vont pouvoir rechercher, trouver et exploiter les données partagées.

Selon les types d'utilisateurs, ils auront des besoins d'accès à la donnée différents

- Le management ainsi que les équipes purement opérationnelles auront généralement besoin de consulter des rapports statiques, leur donnant une information précise, consommable rapidement.
- Les analystes métier ou data préfèreront disposer de moyens de consultation plus libres, au travers de rapports dynamiques voire éditables pour laisser libre court à leur recherche d'informations valorisables.
- Les Data Scientists et Engineers exploiteront directement les données via des accès adHoc SQL et/ou python
- Les développeurs applicatifs, enfin, consommeront plus volontiers des APIs



V. Conduite du changement et adoption

La plupart des équipes Domaines manipulent déjà de la donnée. Sans doute de façon désorganisée, peu maîtrisée. Souvent directement sur leurs propres bases opérationnelles, afin de monitorer leurs applications, de détecter les axes d'améliorations opérationnels, de sortir quelques Kpis simples.

Ici, nous parlons de leur proposer tout autre chose.

Nous les invitons à bord d'un train qui leur permettra de découvrir leur patrimoine de données sous un nouveau jour. Ils exploreront leurs données à travers différentes applications et domaines de l'entreprise.
Ce périple favorisera le partage de la culture Data au sein de l'organisation. Il encouragera également l'enrichissement de leurs données et de leurs idées.

Tout cela sera possible grâce aux insights qu'ils feront ressortir de ce melting pot de datasets. Cette expérience leur offrira une nouvelle perspective sur la valeur de leurs données.

Organiser ce voyage n'est pas une tâche facile. Notre défi est de donner envie à tous ces nouveaux utilisateurs d'embarquer dans cette aventure.
Actuellement, ils évoluent dans un monde familier. Ils gèrent leurs systèmes opérationnels et fournissent des services qu'ils maîtrisent bien. C'est leur zone de confort.
Nous les invitons à quitter ce terrain connu pour explorer un nouveau monde plein de possibilités. Dans ce nouvel environnement, ils pourront valoriser davantage leurs services existants. Ils auront l'opportunité d'améliorer considérablement les processus qui sous-tendent ces services.

Mais ce n'est pas tout. Ce voyage leur ouvrira également des portes vers l'innovation. Ils pourront même envisager la création de services entièrement nouveaux, enrichissant ainsi leur offre.
Ce passage d'un monde à l'autre représente une évolution majeure dans leur approche du travail et de la création de valeur.

Pour y parvenir, il faudra procéder de façon itérative, en commençant par la mise en place de MVP, et trouver le bon panel d'utilisateurs de la première heure.
Il devront être capables de comprendre le projet, la plateforme, les outils et faire évoluer les concepts avec vous, tout en étant promoteur de cette plateforme auprès des autres équipes.

Faites vous accompagner de vos services de communication pour donner des informations régulièrement sur les avancées de la construction de la plateforme.
Faites en sorte d'avoir le meilleur sponsorship au plus haut niveau de management mais aussi des convaincus dans le middle management.
Bâtissez des plans de formation accélérée pour les opérationnels, qui vivront au quotidien ces changements dans leur mission.

A l'arrivée, des équipes devenues maîtresses dans l'art de collecter des données, de les transformer, de les partager, de les analyser, de les **VALORISER** !

Des questions ?

Contactez :

Alexandre MILLET-BASSEZ
Leader de la communauté
alexandre.millet-bassez@ntico.com
03 66 72 80 79