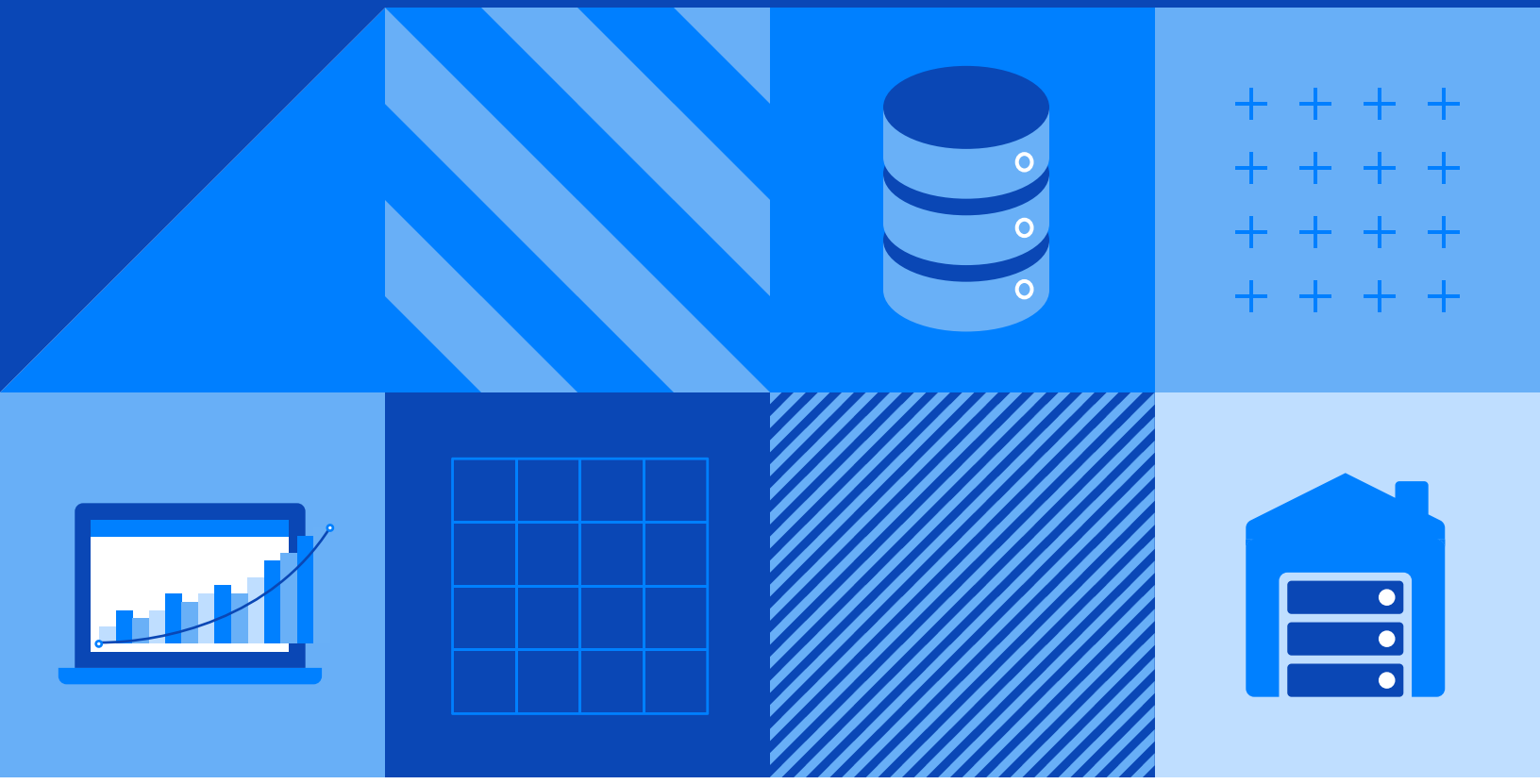
7 conseils pratiques pour la conception de Modern Data Stacks

Les meilleures pratiques de gestion des données que nous avons apprises en concevant des stacks d'Analytics chez J.P. Morgan, Fivetran et bien plus.

Veronica Zhai *et Charles Wang*



Fivetran

Résumé

Ceci est un guide pour construire une Data Stack moderne et découvrir d'autres meilleures pratiques de gestion des données. Vous y lirez les informations suivantes:

autres équipes fonctionnelles au sein de votre entreprise, c'est-à-dire un modèle en étoile. Coordonnez votre action avec celle de vos homologues d'autres équipes qui s'intéressent à l'Analytics. Enfin, établissez quelques indicateurs initiaux pour mesurer la santé de votre entreprise.

5. **Comment gérer votre équipe données comme une équipe de R&D –** Les équipes données fonctionnent mieux en gardant les clients, c'est-à-dire les utilisateurs de leurs données, au premier plan et en incorporant les meilleures pratiques des équipes produit et ingénierie.

6. **La pensée systémique, partie I: Optimiser les flux de travail de Data Integration –** Les données d'entreprise peuvent rapidement devenir compliquées. Efforcez-vous de simplifier et de décomposer les silos autant que possible. Les images sont plus éloquentes que les mots – disposez visuellement le plus grand nombre possible de composants mobiles en un seul endroit.

7. **La pensée systémique, partie II: Utiliser l'information comme levier –** Utilisez les diagrammes de boucles causales pour organiser visuellement les indicateurs et identifier les points de levier dans vos processus commerciaux.

Préface

J'ai commencé ma carrière en tant que trader d'options chez J.P. Morgan et ai ensuite conçu sa première Data Stack moderne. Au début, j'ai été frappée par la sophistication du système financier et la nature quantitative du travail quotidien. Il est orienté sur les données dans un sens très direct et littéral: les traders travaillent avec huit écrans, chacun rempli de centaines de chiffres et d'indicateurs mobiles. Si un trader ne gère pas correctement une option qui expire à un prix d'exercice particulier en fonction de l'évolution des marchés, il peut littéralement gagner ou perdre des centaines de milliers de dollars en quelques minutes. Les données sont de l'argent, et la technologie et les données peuvent être les principaux atouts concurrentiels d'une entreprise.

Malgré le potentiel transformateur des données, les entreprises de toutes tailles ont souvent du mal à les rendre utiles. La mise en place d'une infrastructure de données efficace peut impliquer une multitude de méthodes et d'outils, et la profession des données contient de nombreuses perspectives et approches concurrentes. Les bonnes solutions ne sont pas évidentes, même pour des mastodontes comme J.P. Morgan dont le travail est très sensible et quantitatif par nature.

Actuellement, je suis un leader principal de l'Analytics chez Fivetran, un leader du marché de la Data Integration. J'ai découvert Fivetran lorsque j'ai commencé à faire des recherches sur différentes technologies permettant l'ingénierie des données. Plus j'en ai appris sur Fivetran, plus je suis tombée amoureuse de ce que l'entreprise représente.

Tout au long de ma carrière, j'ai dû apprendre à concevoir des systèmes d'information complexes à la dure. Aujourd'hui, je souhaite vous faire part de certaines des meilleures pratiques que j'ai apprises et qui vous aideront à réussir.

1. Comment concevoir une Modern Data Stack

Une Data Stack est une suite d'outils qui permettent la Data Integration. La Data Stack moderne se compose d'outils de données Cloud native axés sur l'automatisation, la réduction des coûts et la facilité d'utilisation pour les utilisateurs finaux tout au long du cycle de vie de la gestion des données. Avec la croissance du cloud et des plates-formes de données basées sur le cloud tout au long des années 2000, les entreprises peuvent aujourd'hui facilement lancer leurs efforts de Data Integration en utilisant une suite d'outils de données Cloud native.

Les entreprises traditionnelles peuvent être confrontées à un processus lent et pénible pour passer des technologies sur site aux technologies Cloud, avec des coûts fixes élevés et rigides et des problèmes de performance liés aux mégadonnées. Des cycles d'approvisionnement plus lents, des volumes de données plus importants et des risques plus élevés peuvent tous compliquer et ralentir le processus. Les entreprises plus récentes, en revanche, peuvent prendre un nouveau départ avec des outils de données Cloud native.

Dans les deux cas, la conception d'une Data Stack moderne suit ce cadre:

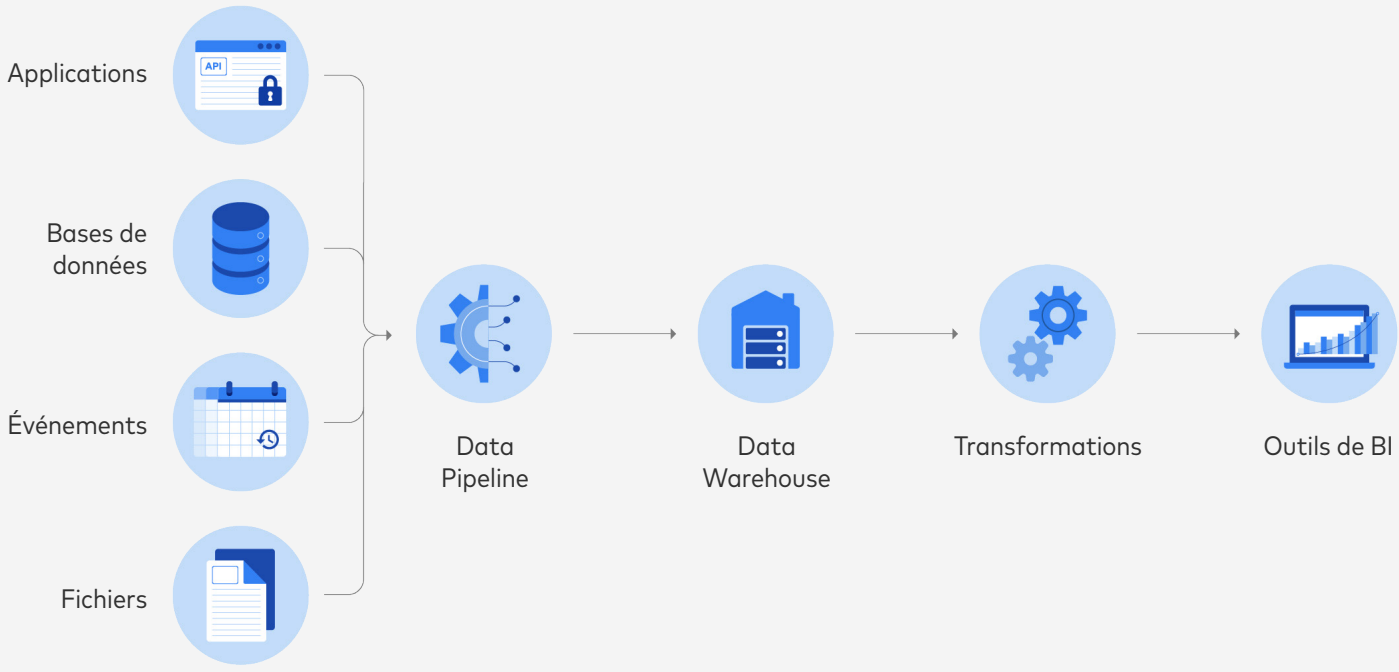
1. **Data Warehouse –** D'abord, configurez un data warehouse basé sur le Cloud. Différents data warehouses offrent une évolutivité, des modèles de tarification, des dialectes SQL et d'autres caractéristiques différents. Parmi les exemples notables de data warehouses Cloud native, citons BigQuery, Snowflake et Redshift.
2. **Business Intelligence** – Connectez ensuite un outil de BI Cloud native à votre data warehouse. Différents outils de BI offrent des niveaux variables de puissance de visualisation, de convivialité, de collaboration et d'autres fonctionnalités. Parmi les exemples notables d'outils de BI figurent Looker, Tableau, Qlik et Mode.

3. **Data Pipeline** – Vous aurez besoin d'un outil pour extraire les données de vos applications et systèmes opérationnels et les charger dans le data warehouse central. Différents vendeurs de Pipelines ont des approches différentes en matière de facilité d'utilisation, de configurabilité, de sécurité et de service client. Les exemples incluent Fivetran, Stitch, Xplenty et Matillion. Les Data Pipelines [1} sont communément désignés par les acronymes « ETL » (Extract-Transform-Load, Extraire-Transformer-Charger) ou « ELT » (Extract-Load-Transform, Extraire-Charger-Transformer), en fonction de la séquence d'actions impliquées dans le déplacement des données d'une source à une destination.

4. **Transformation Data** – Enfin, vous aurez besoin d'outils pour transformer les données en modèles pour le reporting et la modélisation prédictive. De nombreux Data Pipelines et outils de Business Intelligence comprennent des outils de transformation ; Fivetran Transformations en est un exemple.

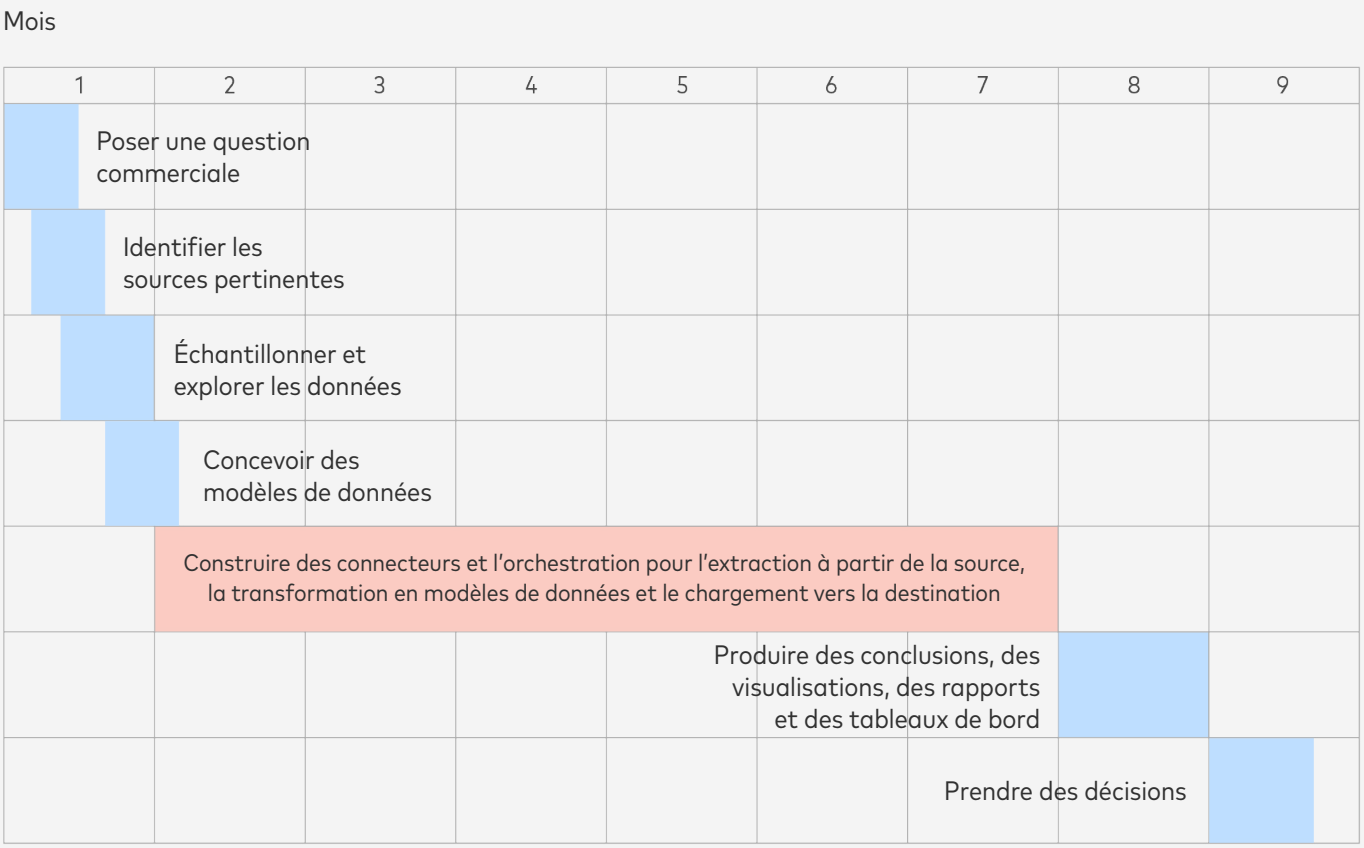
La plupart des vendeurs proposent des essais gratuits. Pour des comparaisons détaillées, consultez des publications sectorielles comme Gartner.

Le flux exact, de gauche à droite, est illustré ci-dessous:



L'ingénierie des données nécessaire pour construire ce flux de travail à partir de zéro peut constituer un obstacle majeur pour les entreprises de toutes tailles. La production et la maintenance peuvent être chronophages et onéreuses, même avec une équipe nombreuse ayant d'importantes ressources.

Les diagrammes de Gantt et les diagrammes de flux de travail peuvent vous montrer où les processus commerciaux provoquent des temps d'arrêt. Par exemple, la nature lourde en ingénierie de l'ETL crée souvent des temps d'arrêt pour les analystes:



AUTODESK.

Prenez Autodesk, par exemple: Lorsque Jesse Pederson, vice-président des plates-formes de données et des insights, a hérité du Data Stack d'Autodesk, la Data Integration était un problème majeur.

« Honnêtement, la Data Stack moderne a vraiment permis de libérer nos équipes pour qu'elles travaillent sur des problèmes présentant plus d'intérêt. Je ne reçois plus d'e-mails disant "Urgent – rupture de Pipelines". Je reçois maintenant des e-mails du genre "Quand est-ce que je peux ajouter mes données ?". »

L'équipe importait des données de Salesforce, SAP, Siebel et des produits d'Autodesk tels qu'AutoCAD, Revit et Maya dans un Data Lake S3.

Pour simplifier le processus, J. Pederson a mis en place un Data Pipeline automatisé et un data warehouse Cloud, et a clairement défini un processus de Data Integration. Aujourd'hui, Autodesk a bifurqué ses Data Pipelines, avec deux voies pour la Data Integration et le stockage des données:

- Si la source est un magasin de données structurées, l'équipe de J. Pederson utilise Fivetran pour l'intégration. Les données structurées sont stockées dans Snowflake.
- Si la source est un magasin de données non structurées (par exemple, des indicateurs d'utilisation des produits et logiciels d'Autodesk), Autodesk utilise AWS Kinesis pour l'intégration de gros volumes dans S3.
- Les données sont répliquées entre les deux référentiels si nécessaire. Les données d'utilisation produit sont transmises à Snowflake pour faciliter la visualisation et l'Analytics, sous réserve des contrôles de confidentialité d'Autodesk, et des instantanés des données Snowflake sont conservés dans S3 à des fins de référence historique et d'apprentissage automatique.

Grâce aux ressources d'ingénierie libérées de la Data Integration, Autodesk a pu mettre en place un système d'alerte précoce pour prédire l'attrition client, ce qui lui a permis de mieux orienter les efforts de l'équipe de support client.

2. Les pièges à éviter lors de la création d'une Modern Data Stack

Dans la section précédente, nous avons abordé les choses à faire pour la conception d'une Data Stack moderne. Il y a aussi un certain nombre de choses importantes à ne pas faire. Ces erreurs peuvent entraver la capacité de votre entreprise à utiliser efficacement les données:

Choisir l'infrastructure traditionnelle au lieu de migrer vers le Cloud. De nombreuses entreprises conservent leur infrastructure existante dans des centres de données physiques coûteux et nécessitant une maintenance importante.

Les Data Stacks sur site présentent les inconvénients suivants:

- Vous devez estimer les coûts du matériel et construire une capacité excédentaire suffisante pour tenir compte de l'utilisation en période de pointe. Cela laisse beaucoup de capacités inemployées en dehors des périodes de pointe, et c'est plus coûteux en général.

- Une installation personnalisée sur site nécessite beaucoup de configuration et de réglage. La configuration et le réglage nécessitent des compétences spécialisées et constituent une barrière à l'entrée très élevée. Cette approche n'est accessible qu'aux grandes entreprises disposant de ressources importantes et, en raison de sa nature lente, laborieuse et généralement difficile, elle n'est sans doute pas une bonne idée, même lorsqu'elle est possible.
- Les performances de votre Data Stack sont en fin de compte limitées par les contraintes posées par votre matériel existant. Cela contraste défavorablement avec une Stack basée sur le Cloud, où des ressources de calcul et de stockage supplémentaires peuvent être activées et désactivées selon les besoins. Il est difficile de faire évoluer un système sur site pour faire face à une activité intermittente élevée, ainsi qu'à une croissance future.

Une infrastructure de données externalisée et basée sur le Cloud offre de nombreux avantages, notamment une meilleure évolutivité, une plus grande facilité d'utilisation, une meilleure accessibilité et un meilleur coût. Il peut simplifier radicalement le flux de travail de votre entreprise.

La migration vers le Cloud peut être un énorme problème en raison de la complexité inhérente des données ainsi que de la nécessité de maintenir les opérations en exécution pendant la migration.

Envisagez de faire appel à un Data Pipeline automatisé, comme Fivetran, pour faciliter le processus. Oldcastle et Copyright nt utilisé la Data Stack moderne pour passer d'une solution sur site à une solution Cloud et continuer à intégrer des données.

Personnaliser vous-même votre Data Pipeline. La Data Integration ne consiste pas seulement à déplacer des enregistrements d'une source vers un emplacement central. Elle implique de sérieux défis d'ingénierie et des considérations de conception, comme la capacité de lire et de mettre à jour les données de manière incrémentielle, la résistance aux pannes, les schémas normalisés, les migrations de schémas, la parallélisation, le pipelining, etc.

En outre, la nature chronophage et laborieuse d'un Data Pipeline « DIY » détourne les ressources d'ingénierie d'autres tâches liées aux produits ou à l'infrastructure et crée des temps d'arrêt pour les analystes.La solution à cet écueil consiste à recourir à l'externalisation et à l'automatisation chaque fois que cela est possible, afin d'éliminer autant que possible la complexité de la tâche.

Choisir l'ETL au lieu de l'ELT. L'approche traditionnelle de la Data Integration, l'ETL, est tellement omniprésente qu'elle est pratiquement synonyme de Data Integration. Malheureusement, l'ETL, avec son couplage étroit entre l'extraction et la transformation et son recours à l'ingénierie personnalisée, est un flux de travail beaucoup plus fragile. Ce n'est plus la meilleure approche pour la plupart des entreprises, et l'ELT, qui permet de charger automatiquement les données dans un état quasi brut avant d'être modélisées par les analystes, est désormais une option plus pratique.

Une discussion plus détaillée sur l'ETL et de l'ELT dépasse le cadre de ce livre blanc, mais le principal avantage offert par l'ELT moderne est l'économie de main-d'œuvre. C'est le résultat des tendances technologiques, notamment de la chute des coûts de stockage, de calcul et de la bande passante du réseau. Fondamentalement, l'ETL conserve les ressources technologiques au détriment de la main-d'œuvre, tandis que l'ELT exploite les capacités technologiques pour économiser la main-d'œuvre.

Manque de discipline dans la gestion des données. Au fil du temps, votre équipe données s'étoffera et vos efforts de Data Integration prendront de l'ampleur. Vous adopterez de nouveaux outils et donnerez à un plus grand nombre de personnes l'accès aux outils permettant de créer des modèles de données, des tableaux de bord et d'autres actifs de données. Vous pivoterez régulièrement, laissant certains de ces actifs de données obsolètes. La croissance introduit le danger de créer des actifs de données qui sont désorganisés et difficiles à trouver ou à interpréter correctement, créant une forme de dette technique.

Vous devez construire des garde-fous en contrôlant et en éliminant périodiquement les actifs de données qui ne sont plus utiles. Comme pour les jardins, la solution à la dette technique liée aux données consiste à élaguer et à conserver de manière organisée. Lorsqu'il s'agit de communiquer clairement des insights exploitables et fondés sur des données, le moins est souvent le mieux.

Ritual

L'histoire de Ritual, une marque de bien-être par abonnement, illustre la fragilité des Data Pipelines « DIY ». Le Data Pipeline personnalisé de Ritual tombait souvent en panne ou prenait du retard. Les Analytics de l'entreprise dépendaient d'instantanés nocturnes que le Pipeline fournissait avec des données obsolètes, ou pas du tout, ce qui se traduisait par des données de rétention manquantes dans le rapport Looker quotidien de l'entreprise. Brett Trani, directeur des données et Analytics, explique:

« Avec les défaillances et les lacunes des données, les gens ont perdu confiance dans les données et allaient chercher leurs propres sources – feuilles de calcul, plates-formes publicitaires, notes aléatoires – et se retrouvaient avec des chiffres différents pour le même indicateur. Il n'y avait pas de source unique de vérité pour la rétention. »

Grâce à un Data Pipeline entièrement automatisé, Ritual a pu créer une source unique de vérité et a connu une réduction de 95 % des problèmes liés au Data Pipeline. Cela a permis de réduire de 75 % le temps consacré aux requêtes et de tripler la productivité de l'équipe données.

3. Comment réaliser les premiers recrutements dans le domaine des données

Un problème important que nous avons rencontré à Fivetran au fur et à mesure de notre croissance était de savoir comment embaucher rapidement des talents. Ayant embauché plus d'une douzaine de professionnels des données talentueux l'année dernière, je souhaite partager un cadre éprouvé qui vous aidera à vous mettre à niveau.

1. **Première sélection: testez les compétences techniques avancées, en particulier des compétences SQL avancées.**
   Vous pouvez utiliser un logiciel de test tel que HackerRank pour accroître l'efficacité de la sélection.

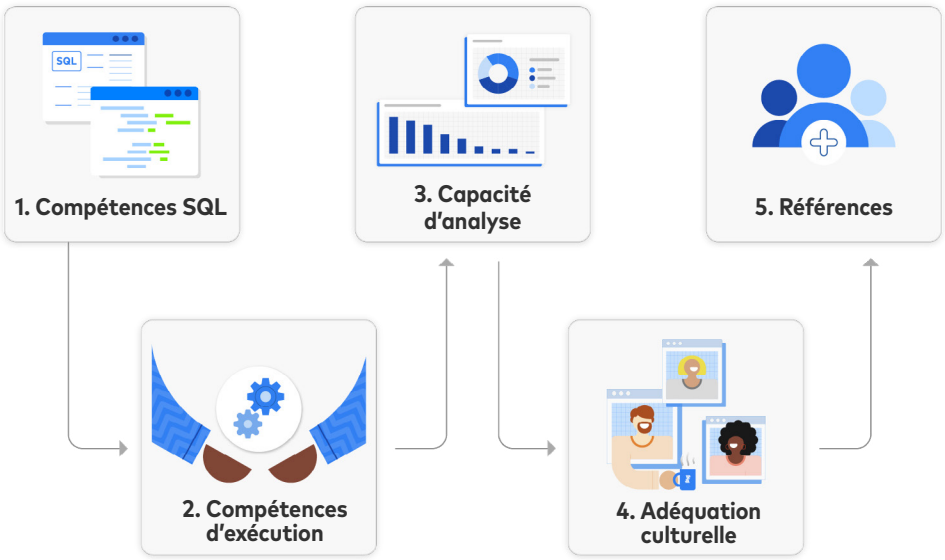
2. **Deuxième sélection: testez les compétences d'exécution.** S'il s'agit de la première embauche, demandez aux candidats de rédiger un plan sur 30-60-90 jours et évaluez leur stratégie commerciale. S'il ne s'agit pas d'une première embauche, demandez-leur de travailler sur du code existant comportant la logique métier. Il s'agit d'un test pratique qui imite les tâches typiques de l'emploi.

3. **Ensuite, vérifiez la capacité d'analyse:** Demandez à vos candidats de présenter des insights et des visualisations à partir d'un échantillon complexe de données, avec les implications pour la stratégie et les décisions commerciales. Vous pouvez également utiliser une étude de cas en temps limité impliquant un cas d'utilisation pertinent pour évaluer leur capacité d'analyse. Cela permettra de tester la capacité de vos candidats à analyser des données et à articuler leurs idées.

4. **Une excellente adéquation culturelle est indispensable:** Nous avons constamment amélioré la densité des talents en recherchant l'adéquation culturelle et l'alignement sur les valeurs de l'entreprise. Donnez la priorité aux nouvelles recrues dont les traits et les capacités complètent votre équipe.

5. **En cas de doute, passez des appels de référence:** Investir une heure dès le départ peut vous éviter de faire une erreur d'embauche coûteuse.

6. **Continuez à embaucher des analystes!** Les outils de données modernes réduisent considérablement les obstacles à la Data Integration et permettent à votre équipe données d'utiliser SQL pour la modélisation et la transformation. Il n'est donc pas nécessaire de recourir à des scripts Python ou Java. Vous pouvez reporter le recrutement de data engineers pendant un certain temps. Au lieu de cela, continuez à constituer une équipe d'analystes. À un moment donné, envisagez d'engager un architecte de données pour optimiser le système global.



Enfin et surtout, misez fortement sur le réseautage et les recommandations pour améliorer la rapidité et la qualité du recrutement. Le fait de mobiliser vos employés pour des recommandations et d'y associer des incitations à la recommandation est très utile.

Vous n'avez peut-être pas les ressources nécessaires pour embaucher des analystes qui répondent à tous les critères énumérés ci-dessus. Ce n'est pas grave – si vous avez des doutes, trouvez des personnes hautement motivées et qui savent s'adapter. Les compétences techniques peuvent être acquises sur le tas.

Les professionnels des données sont actifs dans un certain nombre de communautés de données en ligne et de sites d'emploi. Essayez de contacter des personnes à des adresses telles que:

- Locally Optimistic (locallyoptimistic.com)
- Outer Join (outerjoin.us)

Avec l'aide d'un logiciel, vous pouvez supprimer les tâches fastidieuses de l'ingénierie des données, ce qui permet à vos analystes de se concentrer sur l'Analytics et à vos ingénieurs de se concentrer sur l'amélioration des produits et des infrastructures.

L'automatisation précoce de la Data Integration peut vous aider à établir des priorités lors des premières étapes du recrutement, en vous concentrant davantage sur l'Analytics que sur l'ingénierie, surtout lorsque vous devez démontrer la valeur de votre service en pleine expansion à l'ensemble de l'entreprise.

Prenons l'exemple de Chris Klaczynski, responsable de l'Analytics marketing et utilisateur avancé de la Data Stack modernes chez Databricks.

C. Klaczynski a été la première personne embauchée par Databricks dans le domaine de l'Analytics. Alors que Databricks se développait rapidement, C. Klaczynski a reconnu l'importance d'une solution de Data Integration évolutive:

« Nos nouvelles recrues peuvent se mettre au travail et commencer à concevoir des tableaux de bord immédiatement. Ils n'ont pas besoin de construire des Pipelines ou de se familiariser avec les blocs-notes ou l'écriture de code. Ils peuvent avoir un accès immédiat aux données afin de se concentrer sur les insights et l'établissement de relations avec leurs parties prenantes. »

La Data Integration automatisée s'est immédiatement imposée comme une solution aux besoins de C. Klaczynski en matière Data Pipeline:

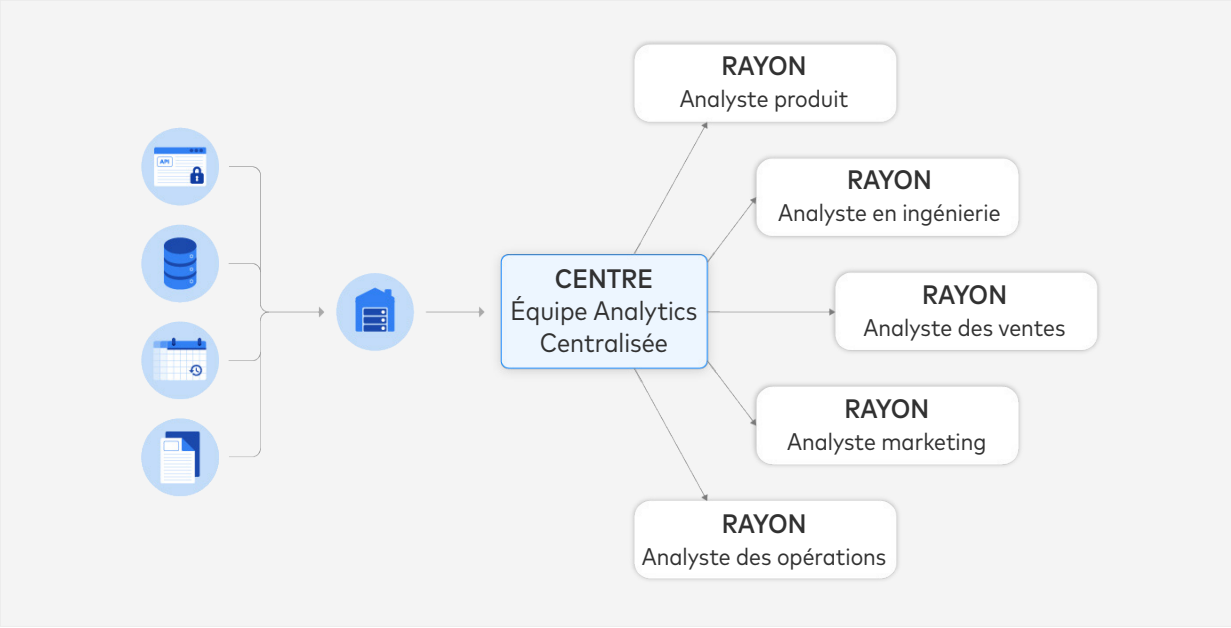
« Au lieu d'embaucher un effectif traditionnel en ingénierie, la Data Integration automatisée nous a permis de nous concentrer sur la valeur commerciale, en embauchant des analystes, des créateurs de tableaux de bord, des personnes expertes en Analytics web et en médias payants. Notre infrastructure est beaucoup plus large et plus avancée qu'il y a un an ou deux. »

Au lieu de construire des Data Pipelines ou des connecteurs Data, Databricks a pu se concentrer sur l'Analytics, la BI et l'architecture de l'information. Les nouvelles embauches ont porté bien plus sur les analystes que sur les data engineers.

4. Que faire au cours des 180 premiers jours

Vos six premiers mois sont cruciaux pour préparer le terrain pour les efforts d'Analytics de votre entreprise. Voici un cadre qui vous aidera à construire une base solide pour un succès rapide.

1. **Concevez une équipe données centralisée:** Une équipe centralisée avec une structure « en étoile » est un modèle supérieur pour la majorité des entreprises, car il permet d'aligner la stratégie et l'exécution. L'équipe Analytics (le centre) doit être directement rattachée au PDG ou à un cadre technique, et les pods (les rayons) spécialisés dans des domaines d'activité particuliers doivent être alignés fonctionnellement avec leurs services respectifs. Ce modèle a bien fonctionné à la fois chez J.P. Morgan, où l'équipe doit soutenir les entreprises à une grande échelle, et chez Fivetran, où l'entreprise a un grand besoin d'évolutivité.



La structure en étoile permet à vos analystes d'établir des relations de travail étroites et d'acquérir une expertise avec des unités fonctionnelles spécifiques de votre entreprise, tout en veillant à ce qu'une équipe centrale d'analystes puisse coordonner et traiter les tâches qui doivent être centralisées.

2. **Travaillez avec des équipes de pairs:** Tout d'abord, identifiez les autres équipes qui utilisent déjà l'Analytics, et la manière dont elles l'utilisent. Créez des alliances en les aidant à automatiser leur Data Integration et à éviter la duplication des tâches. Deuxièmement, déterminez le champ d'action de l'équipe, en identifiant en particulier les tâches qui sont hors du champ d'action afin d'améliorer la concentration et l'exécution.

3. **Alignez les indicateurs fondamentaux avec la direction:** Les dirigeants de votre entreprise doivent s'assurer que la couche BI fait partie intégrante de la stratégie commerciale, car « ce qui est mesuré est géré ». Voici un cadre simple des premiers indicateurs clés de performance importants pour une société SaaS:
   - Indicateurs du chiffre d'affaires
     - Revenu annuel récurrent (ARR)
     - Rétention du revenu net (NRR)
     - Économie d'unité: par exemple, coût d'acquisition du client, efficacité des ventes
   - Ventes et marketing
     - Croissance de la clientèle et taux d'attrition
     - Croissance du revenu d'un mois sur l'autre
     - Indicateurs de conversion et de prospects qualifiés pour le marketing
   - Produit
     - Utilisateurs actifs quotidiens, hebdomadaires et mensuels
     - Parcours client
     - Utilisation des fonctionnalités
     - Net promoter score

Méfiez-vous des indicateurs de vanité qui font un bon effet mais n'influencent pas les résultats qui comptent pour l'entreprise. L'objectif de la définition d'indicateurs est de fournir une orientation claire et d'aligner les motivations de chacun au sein de votre entreprise. Étant donné que les dirigeants de votre entreprise voient plus de pièces du puzzle que les autres personnes de l'entreprise, il leur incombe d'accorder une attention particulière à la définition des indicateurs.

maravai LifeSciences

Maravai LifeSciences est une entreprise leader dans le domaine des sciences de la vie. La direction voulait un aperçu et un reporting plus holistiques sur l'ensemble de Maravai LifeSciences, en commençant par la planification et l'analyse financières (FP&A). Ayant plusieurs filiales autonomes, l'Analytics a été confrontée à des défis tels que:

- Des sources de données disparates nécessitaient un mélange manuel, un nettoyage et une transformation importants pour créer un ensemble de données complet
- Les informations n'étaient pas facilement disponibles et il fallait creuser pour les découvrir
- Pas de système de « source unique de vérité » avec accès à tous les décisionnaires clés

Grâce à un Data Pipeline automatisé, un data warehouse Cloud et un outil de BI Cloud, l'entreprise peut désormais répondre à toutes ses questions financières, notamment:

- Comment est le revenu par rapport à la même période de l'année dernière?
- Quelles sont les ventes quotidiennes moyennes d'une filiale pour le dernier trimestre? Et pour l'année dernière?
- Comment la marge d'une filiale a-t-elle augmenté ou diminué depuis l'année dernière?
- Quels sont les cinq principaux clients de chaque filiale?
- Quelle région client contribue le plus au revenu total?

Maravai LifeSciences prend désormais des décisions commerciales à l'aide de tableaux de bord pour les données Analytics critiques relatives aux finances, aux clients et aux ventes, et prévoit d'ajouter des tableaux de bord pour l'Analytics des produits, des ventes, des inventaires et des profits et pertes.

5. Comment gérer votre équipe données comme une R&D Team

Traditionnellement, une équipe chargée des données est simplement considérée comme une équipe de support, une équipe d'ingénierie ou, de plus en plus, une équipe produit. Cependant, je crois qu'une équipe données est une combinaison des trois.

- **Construisez avec un objectif de produit:** Erik Jones, directeur de l'Analytics chez New Relic, résume brillamment cet aspect: « Une équipe Analytics performante doit également savoir recueillir les exigences, définir la portée, gérer les attentes, le marketing et le déploiement, former les utilisateurs finaux et, finalement, favoriser l'adoption de ce qui est conçu. » L'équipe données devrait se concentrer sur le fait de permettre le libre-service. Je recommande d'utiliser l'adoption de l'utilisation et le NPS comme indicateurs principaux pour l'équipe.

L'utilisation et la satisfaction client sont les indicateurs les plus importants pour une équipe données. L'objectif principal d'une équipe données est d' aider une entreprise à prendre de meilleures décisions. Cela signifie que votre équipe données doit permettre et promouvoir l'utilisation des rapports, des tableaux de bord et des autres actifs. La fréquence d'utilisation des rapports, des tableaux de bord et des autres outils d'Analytics par les membres de votre entreprise est un indicateur clé de performance (KPI) important et un contrôle de santé en soi, car cela démontre que les gens utilisent réellement les actifs de votre équipe. Envisagez de créer un tableau de bord qui mesure l'utilisation de vos autres tableaux de bord. En fin de compte, vous devez vous efforcer d'obtenir l'adoption universelle de la prise de décision fondée sur les données.

- **Opérez avec des principes d'ingénierie:** Une équipe Analytics performante doit également investir au moins 25 % de ses ressources dans la mise en place d'une infrastructure de données facilement navigable et évolutive. L'exploitation de ces principes d'ingénierie améliorera l'efficacité opérationnelle : journaux des demandes des utilisateurs, sprints bihebdomadaires, mise en œuvre de processus de révision du code et d'assurance qualité, automatisation continue et documentation exhaustive.

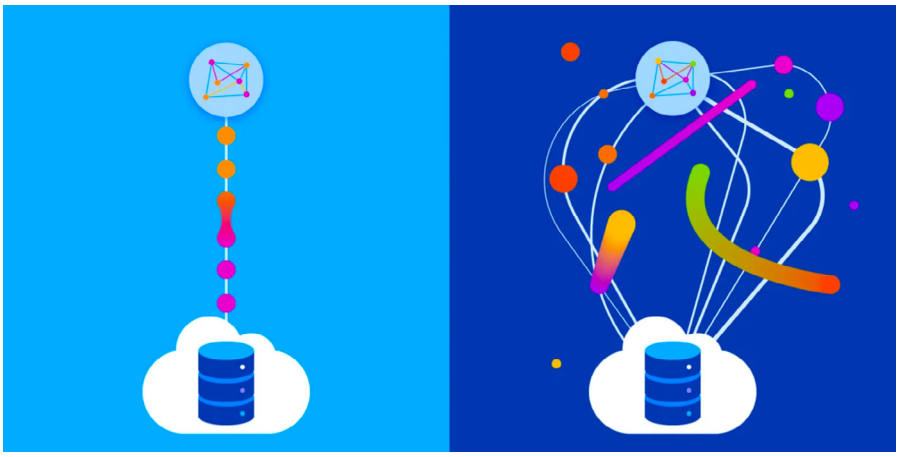
- **Service avec une mentalité centrée sur le client:** L'équipe Analytics doit mettre en place une fonction de succès client technique qui assure l'accueil et l'assistance continue, répond aux problèmes remontés par les utilisateurs, travaille avec les équipes partenaires pour résoudre les problèmes de production et développe des supports de formation pour les utilisateurs finaux. Ces fonctions peuvent être attribuées à des rôles dédiés au succès client technique à mesure que l'équipe s'agrandit.

6. La pensée systémique, partie I: Optimiser les flux de travail de Data Integration

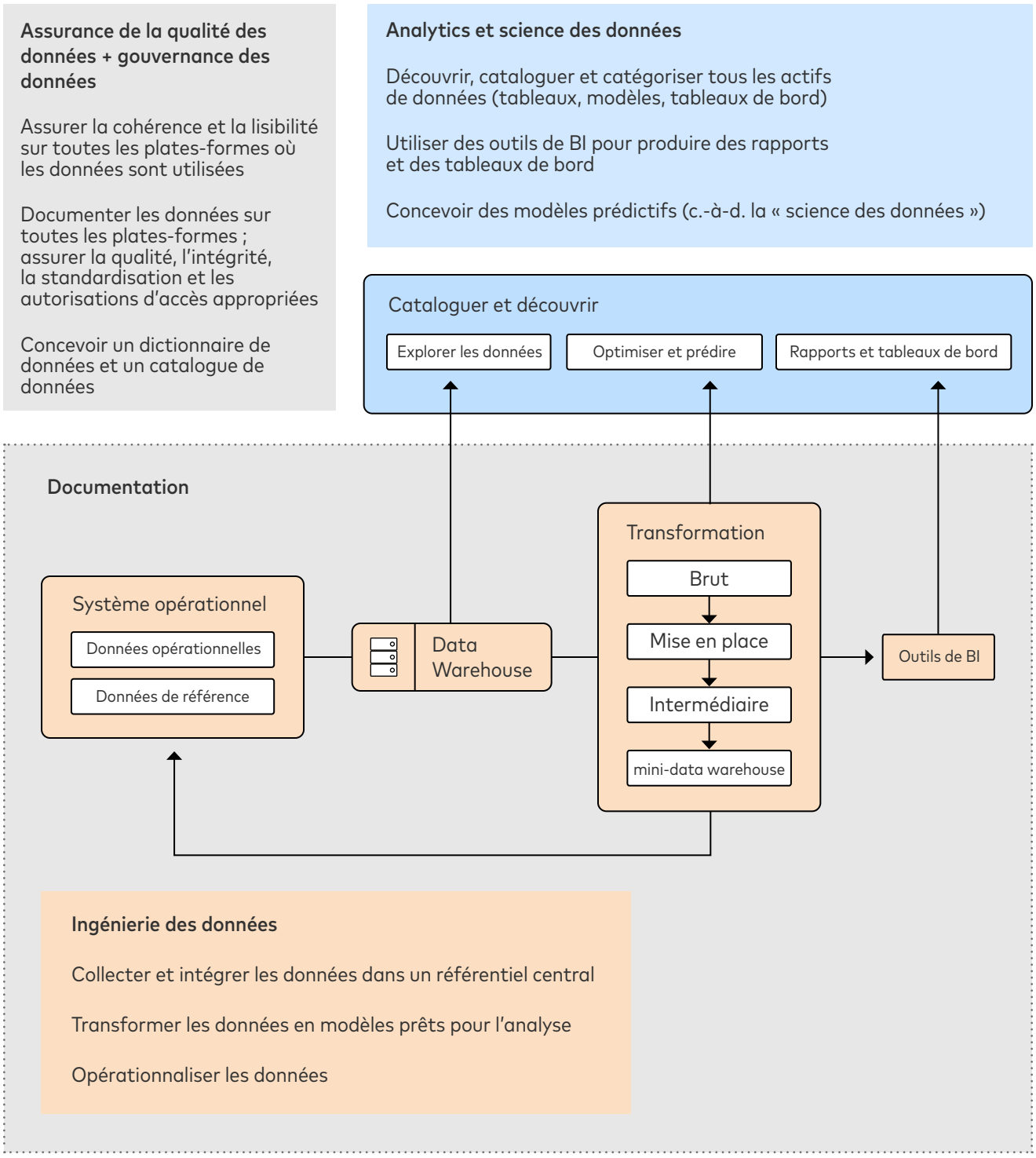
Abordez vos efforts de modernisation des données avec curiosité et patience.Vous devrez acquérir une solide maîtrise des éléments mobiles impliqués dans la Data Integration, identifier les goulets d'étranglement et rechercher les opportunités d'élimination de la complexité chaque fois que possible.

Les données d'entreprise sont complexes et chaotiques

La complexité des données d'entreprise complique également les flux de travail liés à l'intégration de ces données. Comme beaucoup de grandes entreprises, J.P. Morgan a connu de nombreuses fusions et acquisitions au cours du siècle dernier et a dû intégrer de nombreux systèmes différents produisant des données. Cette complexité est amplifiée tout au long du cycle de vie de la gestion des données : travail d'analyse et de rapprochement, intégration des systèmes et Data Integration, surveillance et gouvernance des données, standardisation entre les entreprises, autorisations complexes et problèmes de performance fondamentaux liés au déplacement, à la transformation et à l'interaction avec des pétaoctets de données.



Le diagramme de flux de travail suivant illustre la complexité de la gestion des données d'entreprise tout au long de leur cycle de vie:



La pensée systémique peut vous aider à réussir

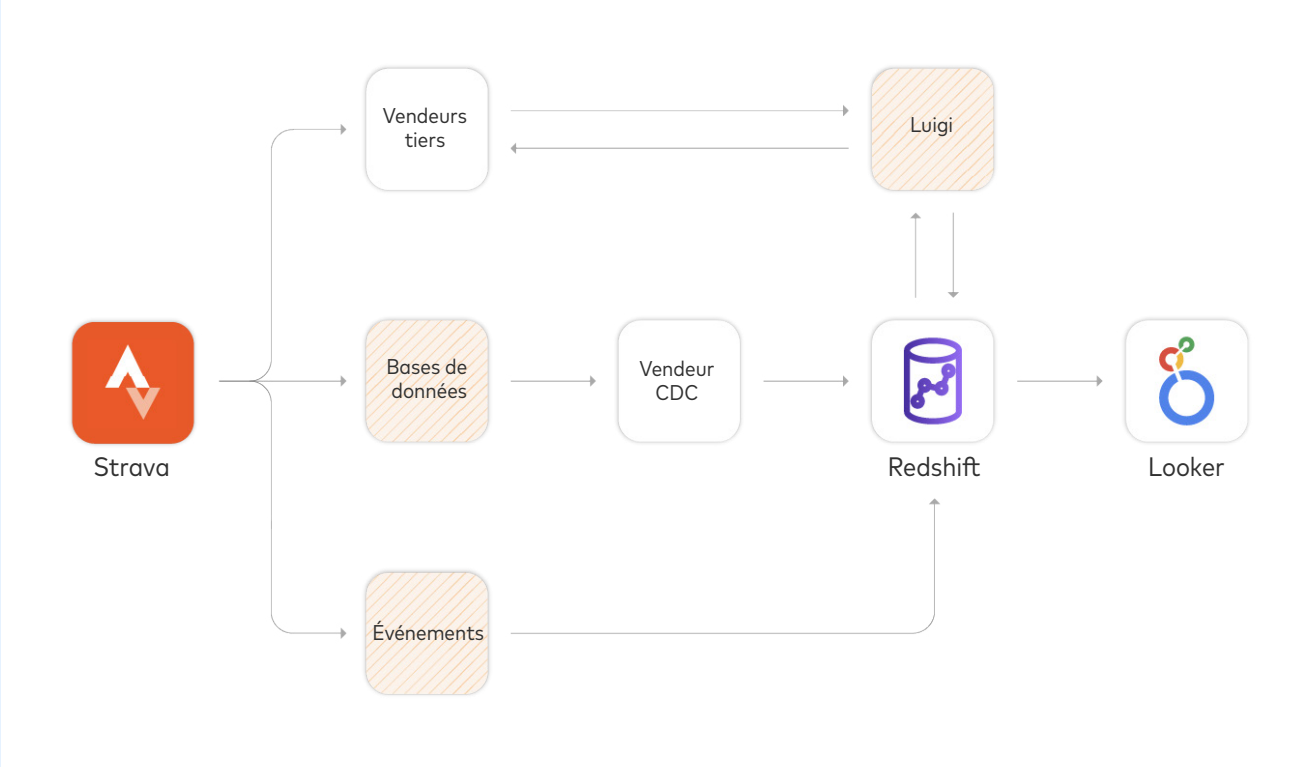
Je suis tombée sur le succès en gérant des données pour le secteur du financement, qui gère des capitaux d'une valeur d'un billion de dollars. Les commerciaux, les traders et d'autres encore faisaient sans cesse des requêtes de données. Cependant, la plupart de ces demandes prenaient du temps à être exécutées, impliquaient de petites optimisations locales et n'avaient qu'un impact limité. J'ai découvert qu'en plaçant stratégiquement des transactions au niveau macro pour faire correspondre les actifs et les passifs, on pouvait libérer des centaines de millions de dollars de capital, ce qui augmentait l'avantage concurrentiel de l'entreprise. Ces actions dépassaient le mandat des équipes individuelles, car aucune d'entre elles ne pouvait accéder à toutes les données.

Notre première tâche a donc été de décomposer les silos et de centraliser les données. Nous avons ensuite exploré les algorithmes d'apprentissage automatique pour analyser comment allouer le capital aux clients afin de maximiser le retour sur investissement. Étant donné qu'il était extrêmement complexe d'automatiser la production de tout indicateur au niveau du système dans une grande entreprise, le fait de fournir une visibilité en temps réel de l'indicateur macro et de l'optimiser à l'échelle globale s'est avéré avoir une valeur illimitée.

L'équipe de Daniel Huang, responsable de l'ingénierie des données chez Strava, peinait à tenir le rythme pour répondre aux besoins de l'entreprise de 12 ans à la croissance rapide.

Finalement, les membres ont été contraints de réfléchir à l'avenir de la culture de l'ingénierie des données de leur entreprise. « Cela a commencé par le passage à une plate-forme », se souvient D. Huang. « Notre rôle en tant que data engineers devrait être de concevoir la plate-forme et de guider les gens dans l'utilisation de celle-ci. Laisser la plate-forme répondre aux besoins en matière de données. »

L'infrastructure de données d'origine de Strava est représentée ci-dessous:

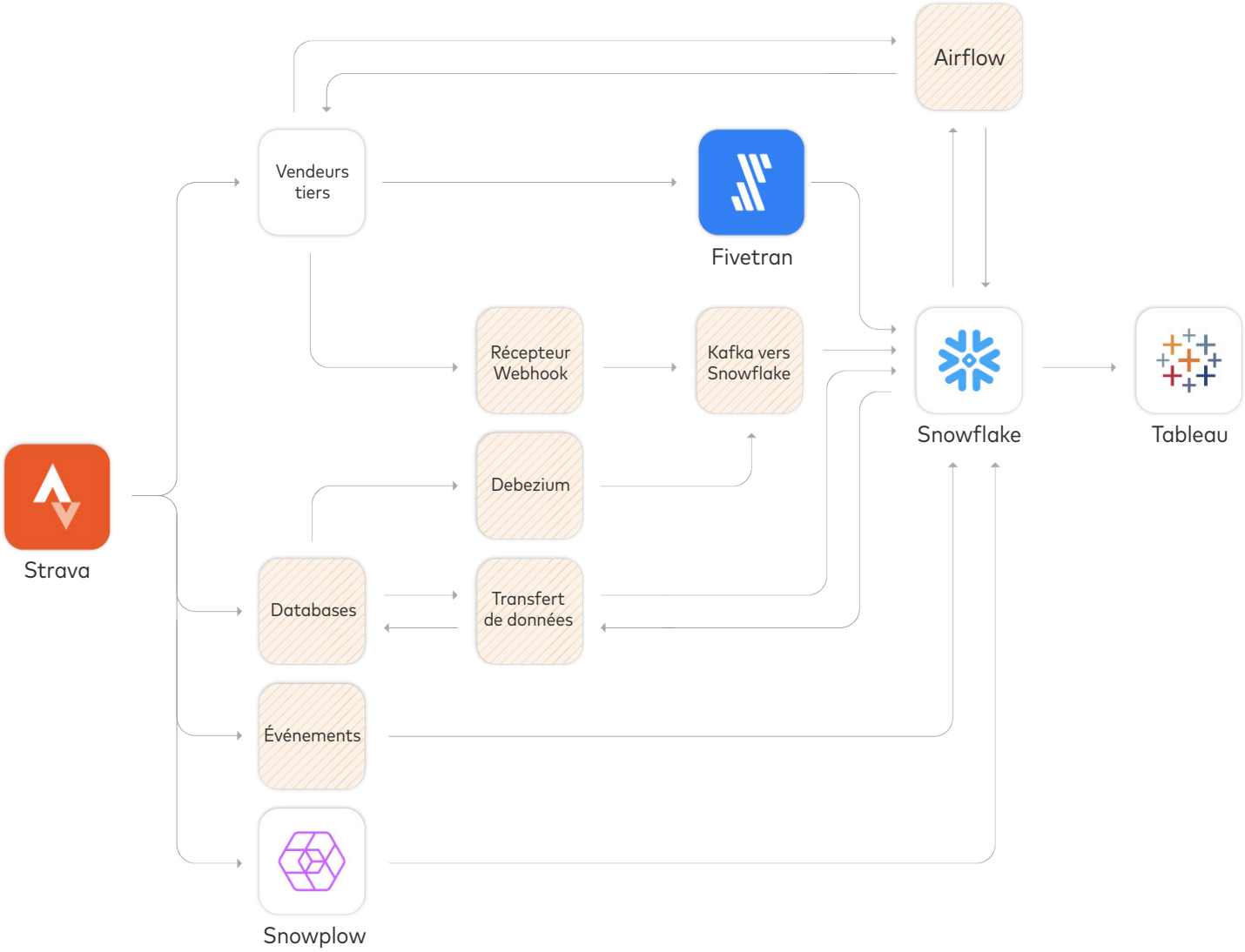


Selon Dimensional Research, 63 % des entreprises recourent encore à des scripts manuels, alors même que les entreprises brassent plus de données, plus rapidement que jamais. En fait, 72 % des entreprises ont désormais besoin que les données soient déplacées quotidiennement, toutes les heures, voire toutes les quelques secondes.

« Au début, toutes nos tâches d'ETL étaient rédigées par quelques data engineers », explique D. Huang, « ce qui signifie que nous devions assurer la maintenance de toutes ces tâches. Nous devions réparer ces tâches au lieu de concevoir les infrastructures ou les services sous-jacents. En résumé: nous devenions l'interface pour les données plus souvent que nous ne le voulions, et nous devenions un goulet d'étranglement pour l'entreprise. »

Avec une nouvelle vision, Strava a mis en place une Data Stack Cloud reposant sur Snowflake en tant que data warehouse, Tableau en tant qu'outil de BI et Fivetran en tant que fournisseur de Data Pipeline pour faire parvenir automatiquement les données dans Snowflake à partir de vendeurs tiers.

Ce qui suit représente l'approche moderne de Strava en matière de gestion des données, en utilisant la Data Stack moderne:

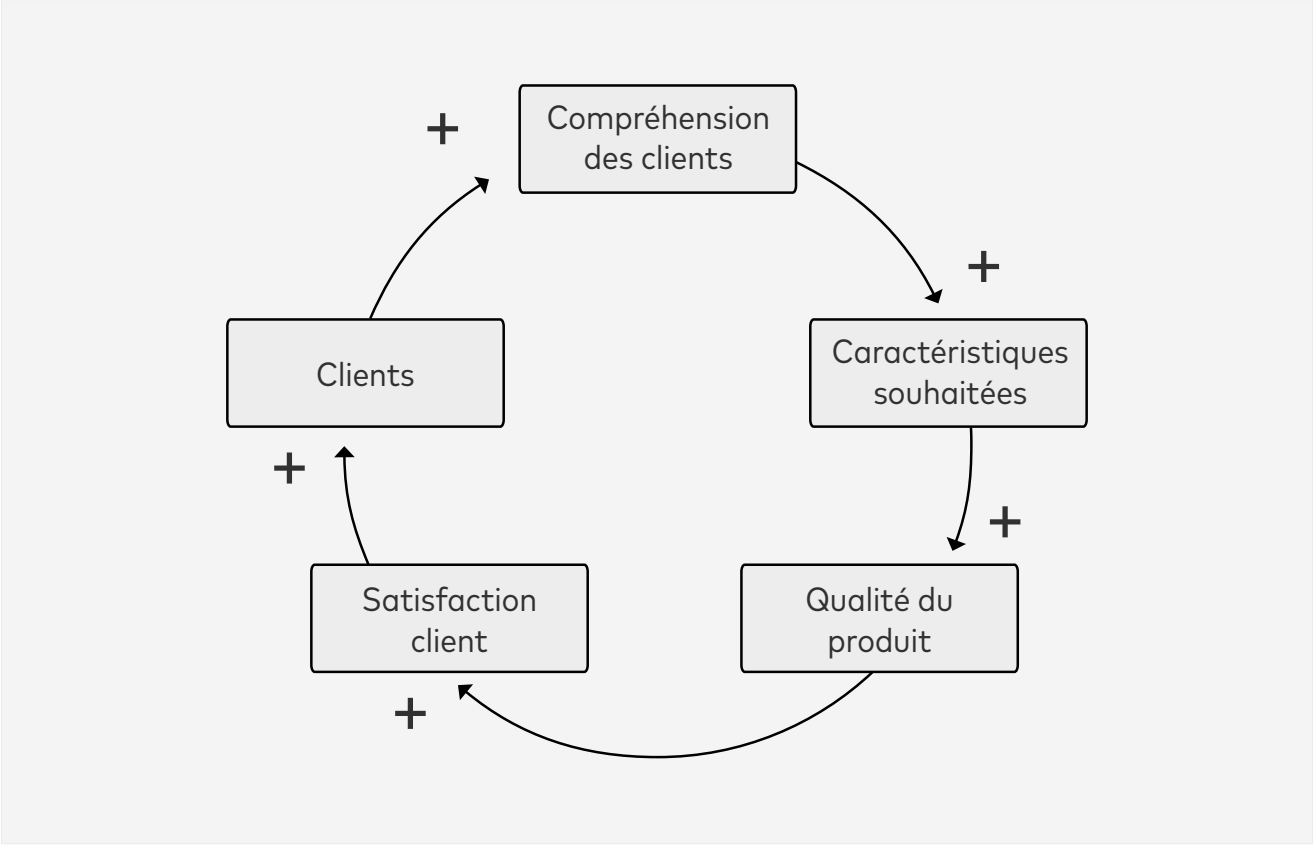


« Nous sommes encore une petite équipe, mais nous avons fait des progrès », déclare D. Huang, en réfléchissant aujourd'hui à mi-parcours. « La facilité d'utilisation des outils basés sur le Cloud a libéré notre équipe pour réfléchir à la culture des données de l'entreprise dans son ensemble, et nous avons mis en place une catégorisation des utilisateurs de données internes pour mieux comprendre et répondre à leurs besoins. »

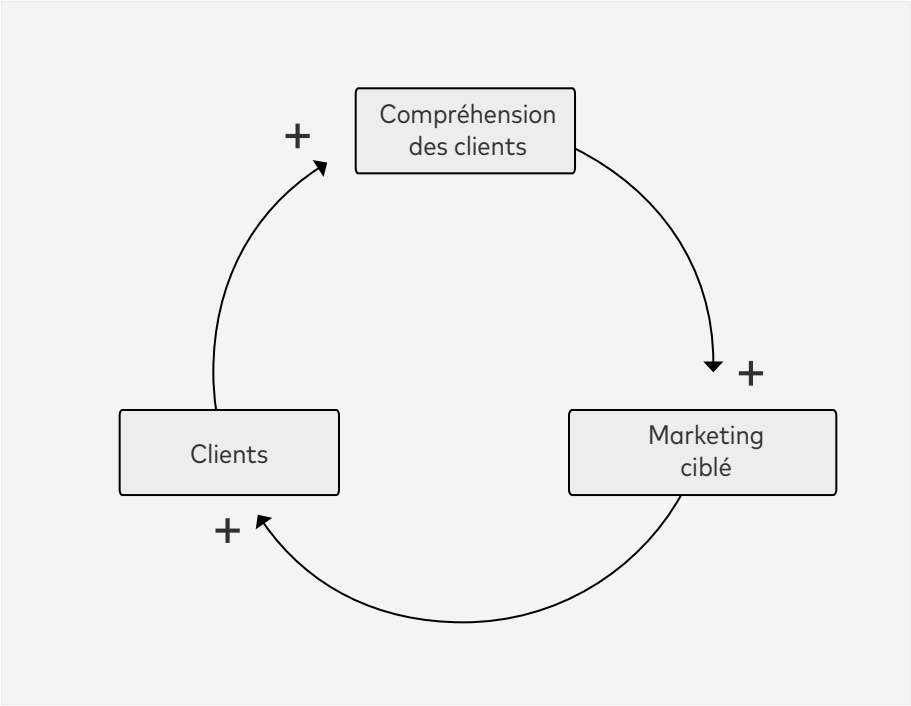
7. La pensée systémique, partie II: Utiliser l'information comme levier

Une fois que vous avez conçu une Data Stack qui fonctionne correctement, vous pouvez identifier les cycles vertueux (et vicieux) qui ont un impact sur vos activités. Pour prendre des décisions efficaces, il est essentiel que les analystes comme les parties prenantes comprennent le modèle économique, les contraintes et les motivations de l'entreprise.

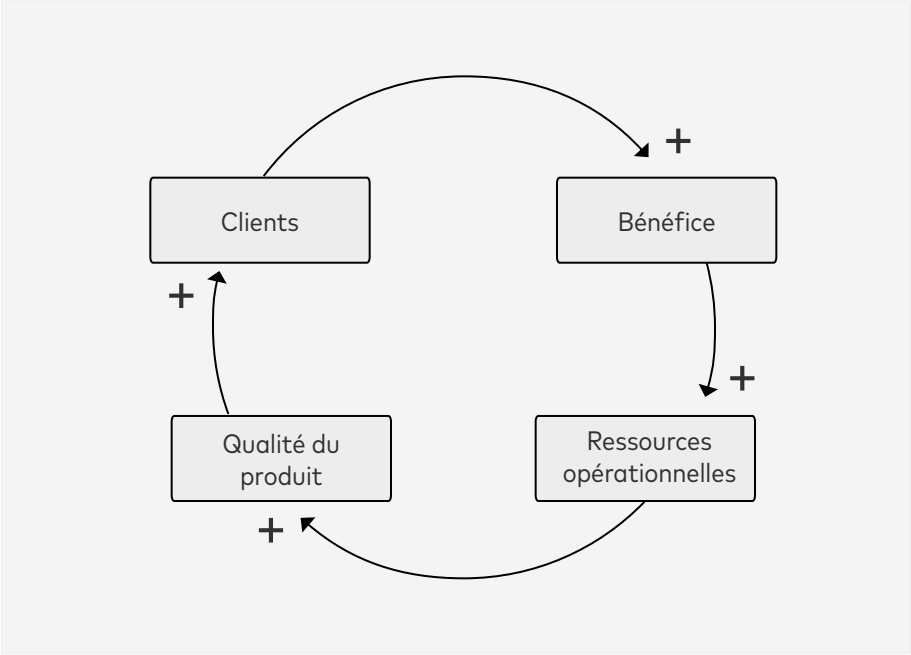
Identifiez les « volants d'inertie » qui font le succès de votre entreprise. Par exemple, une bonne compréhension de vos clients vous permettra d'obtenir un produit de qualité supérieure, ce qui favorisera la satisfaction client et augmentera votre clientèle:



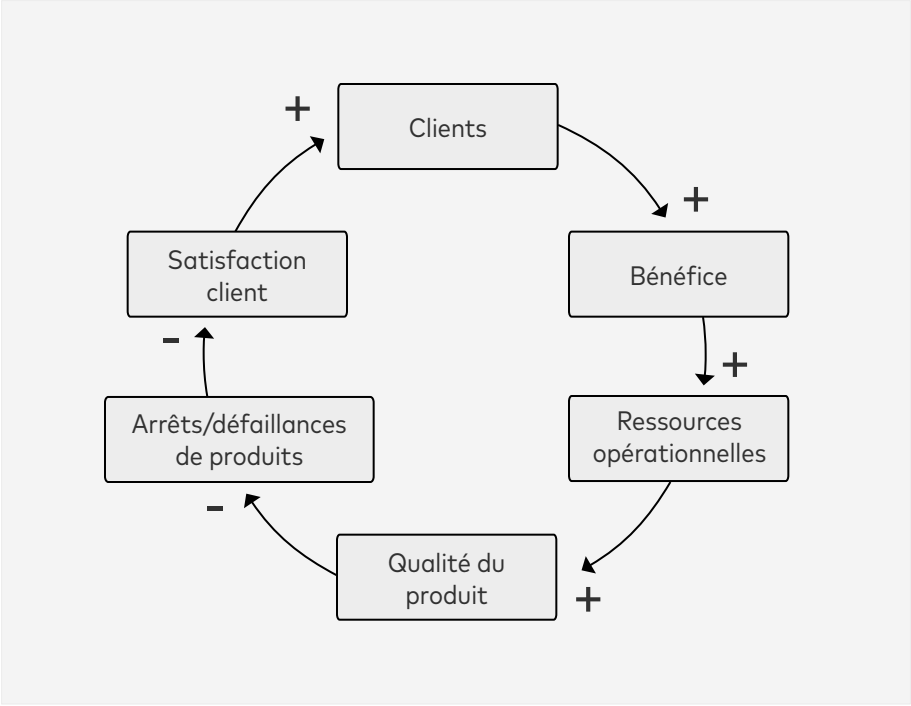
Sans parler de l'amélioration de vos efforts de marketing:



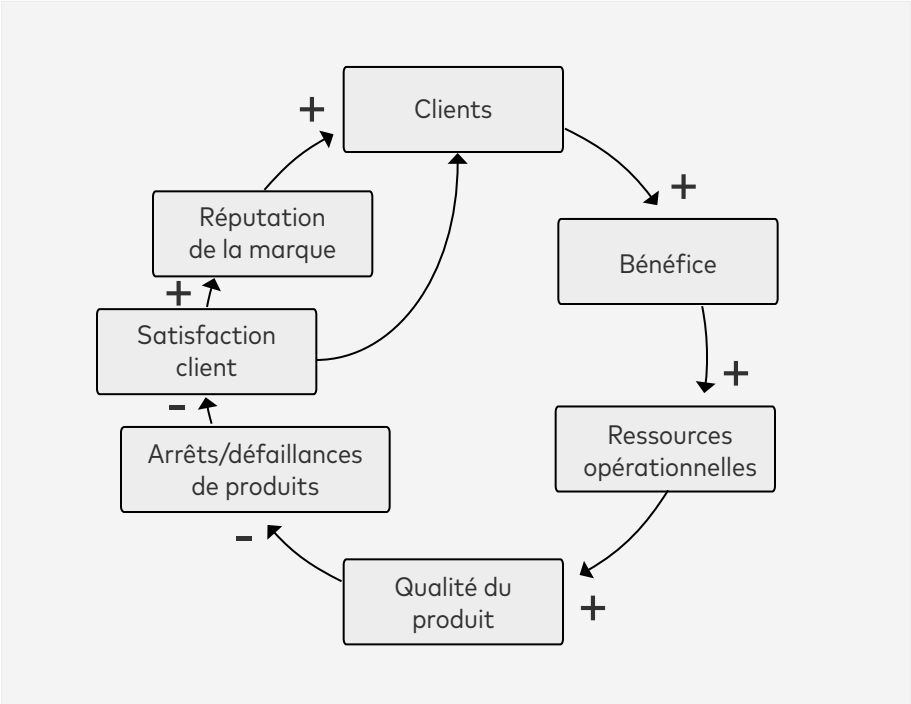
Plus votre produit est bon, plus il y a de clients, de bénéfices et de ressources disponibles pour l'améliorer encore:



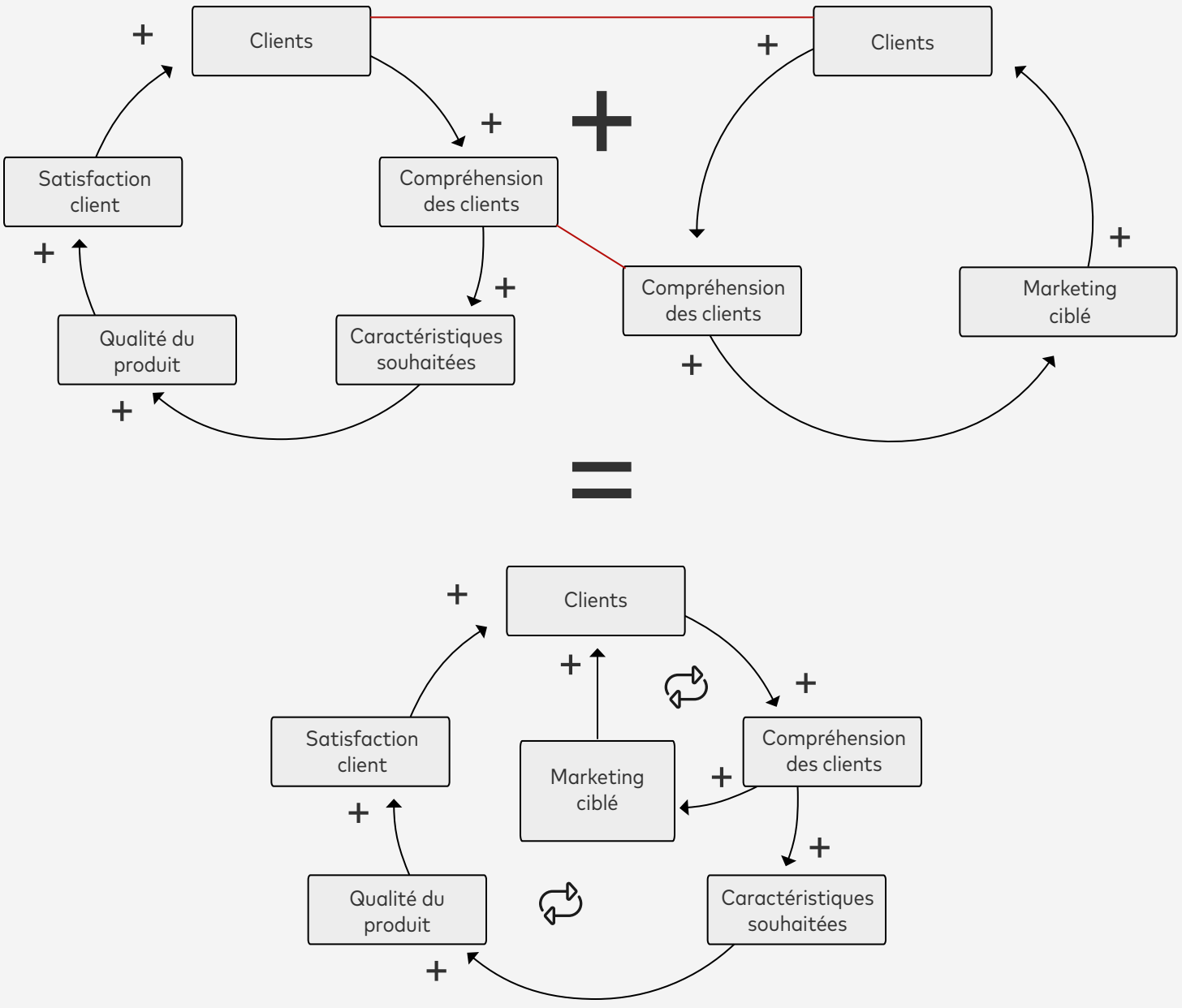
En revanche, les arrêts et les défaillances de produits ont une relation inverse avec la satisfaction client:

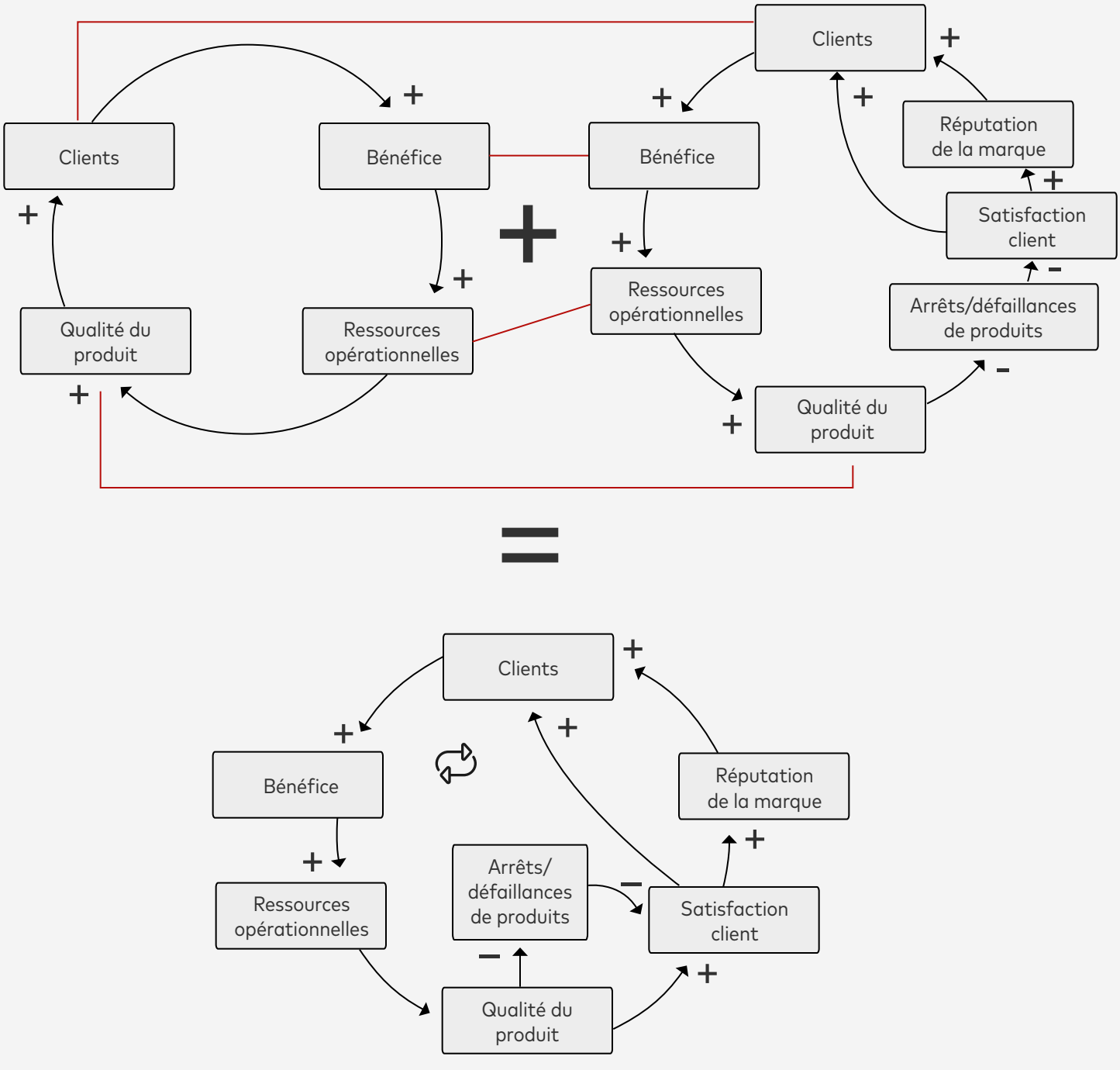


L'effet global est le même lorsqu'il passe par la réputation de la marque:

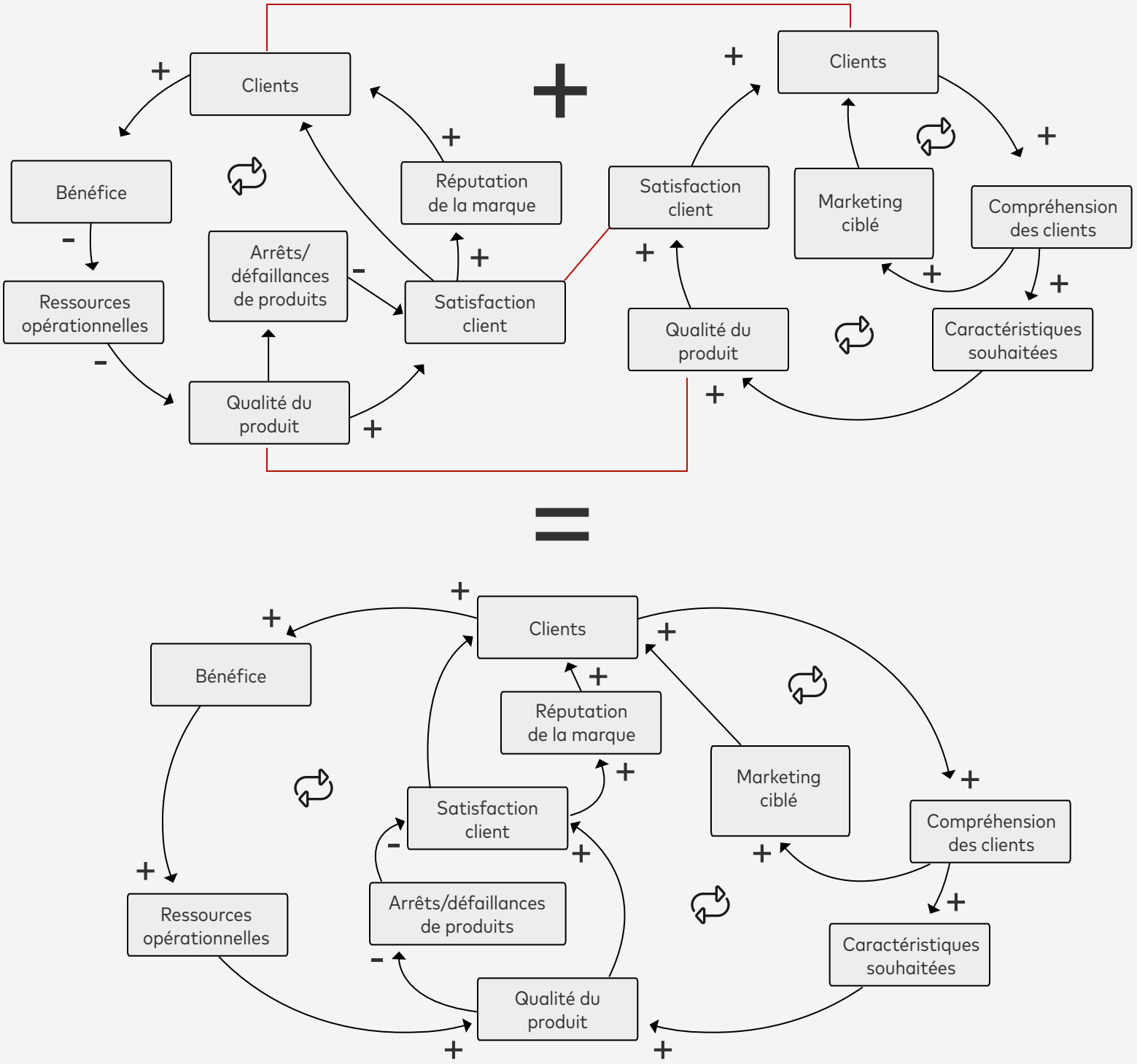


Vous pouvez assembler ces volants d'inertie dans des **diagrammes de boucles causales** pour mieux illustrer les opérations de votre entreprise. Ces volants ont des éléments en commun, ce qui vous permet de les connecter dans des diagrammes de boucles plus larges:




Clients
Bénéfice
Ressources opérationnelles
Qualité du produit
Clients
Bénéfice
Ressources opérationnelles
Qualité du produit
Réputation de la marque
Satisfaction client
Arrêts/défaillances de produits
Clients
Bénéfice
Ressources opérationnelles
Arrêts/ défaillances de produits
Qualité du produit
Satisfaction client
Réputation de la marque

The more comprehensive your diagram, the more you can see all of the moving parts in one place:



Les diagrammes de boucles causales vous permettent de présenter en un seul endroit les composants mobiles qui sous-tendent votre entreprise et ses activités. Plus précisément, vous serez en mesure d'identifier les domaines dans lesquels vous pouvez faire la différence sans essayer de résoudre simultanément tous les problèmes.

Grâce aux données, vous pouvez quantifier les différents éléments de vos opérations. Avec les données et la pensée systémique, vous serez en mesure d'identifier les points de levier qui vous permettront d'utiliser au mieux vos ressources pour promouvoir les bons résultats.

En tant que professionnelle des données, j'aime concevoir un système d'information élégant et je suis convaincue qu'une bonne utilisation des données peut avoir un impact profondément positif sur le monde et même aider l'humanité à atteindre un niveau de conscience supérieur. Si de nombreuses personnes peuvent se sentir triomphantes face aux avancées technologiques significatives réalisées jusqu'à présent dans la gestion des données, je pense que ce n'est que le début d'une révolution des données.

À propos de Fivetran: Développée pour répondre aux besoins réels des data analysts, la technologie Fivetran offre la solution la plus intelligente et la plus rapide pour répliquer vos applications, bases de données, événements et fichiers dans un Cloud Warehouse hautes performances. Les connecteurs Fivetran sont déployés en quelques minutes, ne nécessitent aucune maintenance et s'adaptent automatiquement aux changements de source, éliminant ainsi les tâches d'ingénierie afin que votre équipe données puisse se concentrer sur le développement des insights. Pour en savoir plus, consultez Fivetran.com.