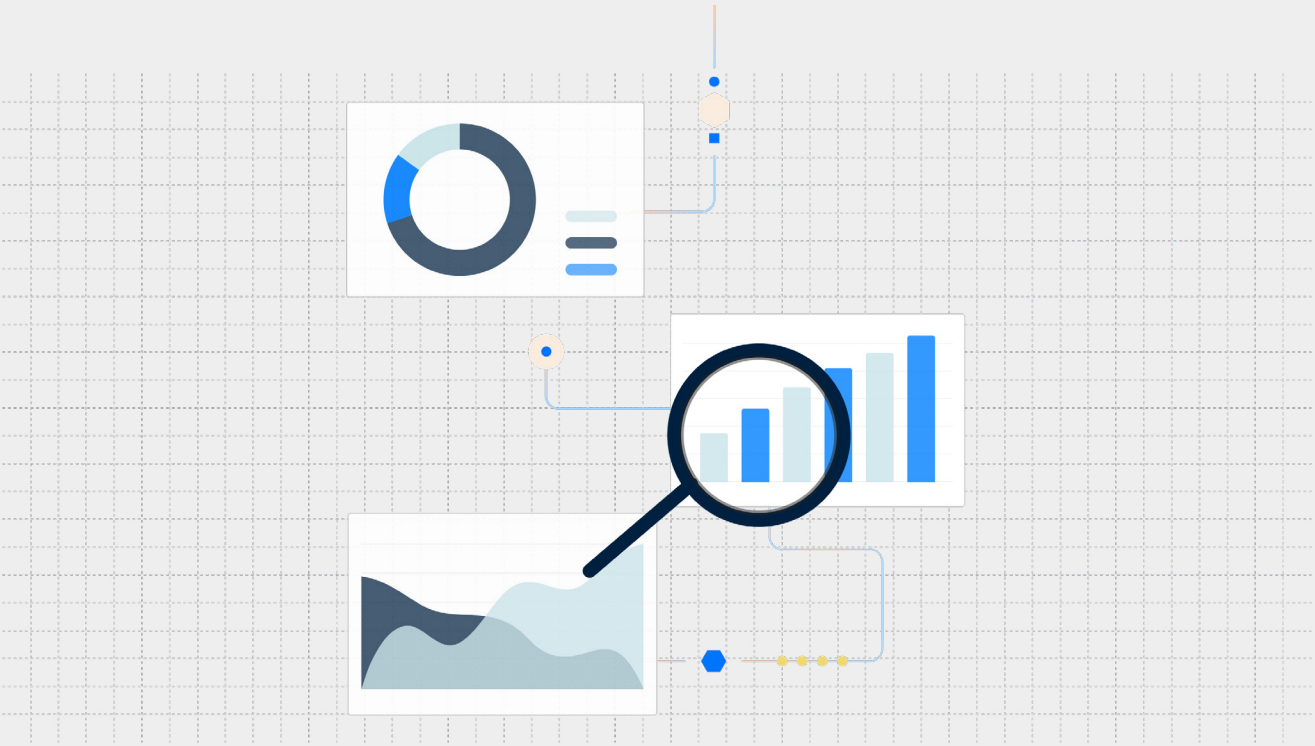
Guide essentiel de la Data Integration

Modèles mentaux pour la Data Integration, l'Analytics et la Modern Data Stack



Fivetran

Sommaire

Introduction

Le monde est inondé de données. Presque toutes les activités laissent derrière elles une empreinte numérique. En tant qu'élément d'une image beaucoup plus vaste, cette trace ouvre l'accès à une connaissance plus approfondie. L'analyse des données est essentielle à tous les secteurs d'activité, car elle aide les chefs d'entreprise à prendre des décisions plus avisées sur les produits à proposer, les segments de marché à cibler, les modalités logistiques, etc. L'essor du cloud computing, de la technologie SaaS et de la **Modern Data Stack** a permis aux entreprises de toutes tailles et de tous budgets d'accéder à des informations qui n'étaient auparavant accessibles qu'aux plus aisées et aux équipes les plus étoffées. Le moment est venu d'adopter une nouvelle approche et un nouvel état d'esprit en matière d'Analytics et de Data Integration.

Sans Data Integration, les données sont cloisonnées, de même que les informations, la connaissance et les insights qui peuvent en être extraits. Les données cloisonnées entravent la collaboration en générant des interprétations partielles et contradictoires de la réalité, ce qui entrave l'établissement d'une source unique de vérité et la mise au diapason de tous les membres d'une organisation.

La Data Integration est également essentielle pour l'intégrité des données, c'est-à-dire pour garantir leur exactitude, leur exhaustivité et leur cohérence. L'intégrité des données désigne la capacité à comprendre parfaitement les données, à avoir confiance en leur exhaustivité et à garantir leur accessibilité en cas de besoin.

Si vous êtes un analyste, un data engineer ou un data scientist (ou que vous supervisez ce type d'experts) dans une organisation qui utilise des systèmes, des applications et d'autres outils opérationnels qui produisent des données numériques, ce guide est fait pour vous. Votre rôle doit vous permettre d'influencer ou de déterminer les outils utilisés par votre entreprise. Plus important, vous devez exercer une influence sur la façon dont les équipes envisagent la résolution de problèmes, la recherche d'insights et la prise de décisions à partir des données. Ce guide porte autant sur l'instauration d'un état d'esprit moderne à l'égard des données que sur l'implémentation de nouveaux outils et processus.

En adoptant un état d'esprit moderne à l'égard des données, vous serez en mesure de tirer parti de l'automatisation pour réaliser des économies substantielles de temps, de talents et d'argent, et d'extraire davantage de valeur de l'utilisation des données. En revanche, tant que votre organisation restera attachée à des méthodes et des mentalités dépassées en matière de Data Integration, vous gaspillerez de l'argent, gâcherez les efforts de vos professionnels des données et perdrez du terrain face à des concurrents plus avisés sur le marché.

Après avoir décrit et évalué les différentes approches de la Data Integration actuellement disponibles, nous vous montrerons comment l'introduire dans votre organisation. Enfin, nous vous aiderons à développer un état d'esprit moderne à l'égard des données et nous vous expliquerons comment apprendre aux autres à adopter le même état d'esprit.

Chapitre 1 : Pourquoi les données sont importantes

La Data Integration désigne les processus utilisés pour gérer et combiner les flux de données provenant de diverses sources. Si toutes les données de votre organisation sont rassemblées dans un seul environnement, vous êtes en mesure d'obtenir une vue complète de vos opérations commerciales. Un des exemples les plus courants est celui de l'examen des parcours des clients en retraçant la manière dont ils interagissent avec le marketing, les ventes, les produits, l'assistance client, etc. Le fait de s'appuyer sur les données pour résoudre des problèmes est une pratique appelée Analytics.

Les cas d'utilisation types de l'Analytics comprennent l'exploration de données pour obtenir des insights afin d'apporter des améliorations dans les domaines suivants :

1. Expériences client
2. Processus et opérations interness
3. Produits, fonctionnalités et services

Les données constituent un pilier essentiel des produits de machine learning ou d'intelligence artificielle. Elles peuvent également être elles-mêmes le produit lorsqu'elles sont présentées de manière lisible et utile aux clients.

Plus vous progresserez dans l'intégration et l'analyse des données, plus vous serez en mesure de promouvoir une culture des données généralisée pour une organisation plus agile, dynamique et innovante.

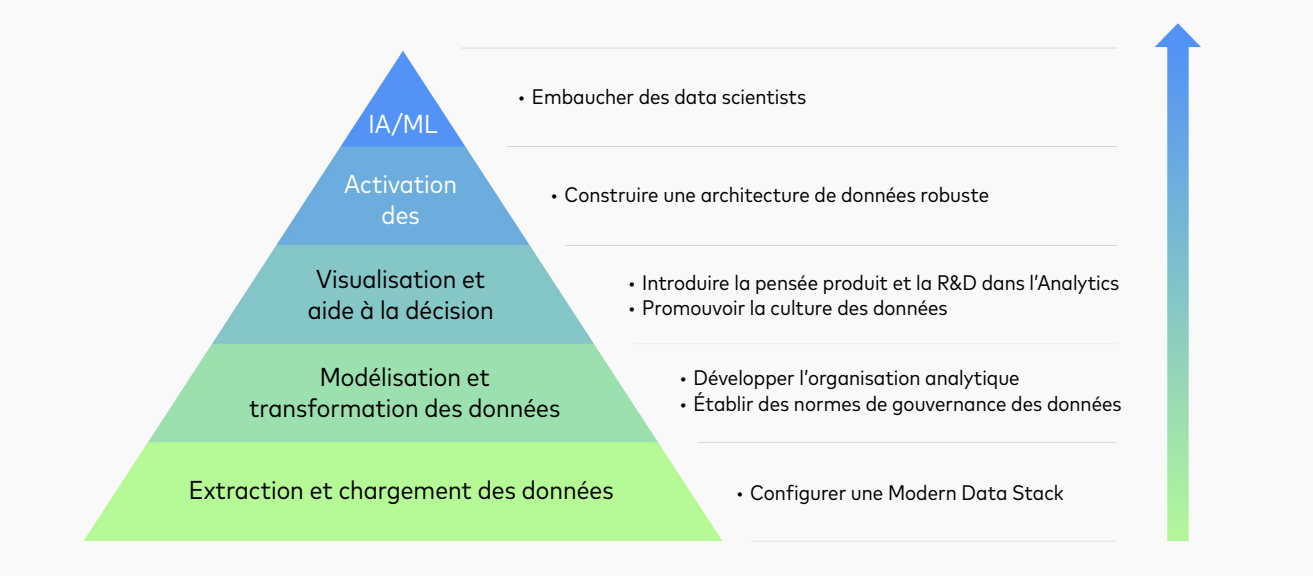
La hiérarchie des besoins en matière de données

L'élaboration d'une approche moderne de l'intégration et de l'analyse des données doit reposer sur une base solide.

On peut placer l'intégration et l'analyse des données dans une **hiérarchie des besoins** avec les étapes correspondantes permettant d'y répondre :

1. **Extraction et chargement des données.** C'est la première étape de la Data Integration. Vous devez rassembler les données et les rendre disponibles sur une plate-forme unique. La meilleure façon d'y parvenir est d'utiliser une série d'outils appelée « Modern Data Stack ».

2. **Modélisation et transformation des données.** Une fois les données provenant de multiples sources rassemblées en un seul endroit, vos analystes peuvent commencer à les transférer dans des structures compatibles avec les tableaux de bord, les visualisations, les rapports et les modèles prédictifs. Au fil de l'évolution de vos besoins, vous devrez étoffer votre équipe chargée des données et établir des normes pour la gouvernance des données.
3. **Visualisation et aide à la décision.** Vous êtes maintenant prêt pour l'Analytics. Vos modèles de données vous permettront de créer une vue d'ensemble de vos opérations. Créez des rapports et des tableaux de bord selon vos besoins. Pour que ces activités se déroulent au mieux, il est indispensable d'intégrer les bonnes pratiques de gestion produit dans la génération des éléments de données et de promouvoir la culture des données dans votre organisation.
4. **Activation des données.** Les données analytiques peuvent être réinjectées dans vos systèmes opérationnels afin de donner aux membres de l'équipe une visibilité en temps réel sur les opérations ou d'automatiser les processus métier qui dépendent d'entrées de données spécifiques. L'intégration de données directement dans les systèmes de production peut nécessiter des outils spécialisés ou une expertise en ingénierie des données.
5. **IA et machine learning.** La construction de systèmes qui tirent parti de la modélisation prédictive se trouve au sommet de la science des données. Il existe d'innombrables applications de machine learning, allant de la réalisation de prédictions sur la base d'une régression linéaire aux véhicules autonomes. À ce stade, votre organisation doit commencer à embaucher des spécialistes comme des data scientists et des ingénieurs en machine learning.



Les avantages commerciaux de l'Analytics

L'Analytics ne consiste pas seulement à faire émerger des informations et des insights pertinents. Au-delà de la compréhension théorique d'une situation, elle éclaire des décisions commerciales concrètes et apporte ce faisant des avantages réels à chaque équipe ou fonction organisationnelle.

L'analyse des données a longtemps été considérée comme un outil de base pour résoudre les problèmes. Si cette pratique perdure aujourd'hui, la modélisation prédictive et à grande échelle occupe une place croissante dans les outils, les technologies et les équipes qui effectuent ce travail, avec notamment l'intelligence artificielle et le machine learning.

L'automatisation et le calcul font bien plus qu'augmenter la vitesse et l'efficacité ou économiser du temps, des talents et de l'argent. L'aptitude à traiter plus de données en quelques minutes qu'un être humain ne pourrait le faire en une vie entière et de tirer des conclusions avisées impossibles à établir autrement introduit une différence qualitative dans les capacités. Cette façon de concevoir l'analyse des données marque le passage à un état d'esprit moderne, où les améliorations des processus et de la capacité de prise de décision sont fortement influencées par les progrès technologiques rendus possibles par une Modern Data Stack.

Citons quelques-unes des applications quotidiennes courantes de l'intelligence artificielle et du machine learning :

1. **Recommandations :** expériences personnalisées dans la publicité, les fils d'actualité sur les réseaux sociaux, les services de streaming et les avis.
2. **Détection des anomalies :** détection de la fraude, reconnaissance d'images et diagnostics médicaux.
3. **Prédictions numériques :** étude des relations de cause à effet, comme le rendement des cultures selon l'utilisation d'engrais, la pression sanguine en fonction des médicaments prescrits ou le chiffre d'affaires par rapport aux dépenses publicitaires.
4. **Agents intelligents :** conception de machines réelles ou virtuelles capables de prendre des décisions de manière indépendante, comme les chatbots, les véhicules autonomes et les chaînes de production automatisées.

Chapitre 2 : Qu'est-ce que la Data Integration ?

La Data Integration englobe l'extraction, le chargement et la transformation des données en modèles de données exploitables. La Data Integration permet d'effectuer des analyses en combinant des données provenant de l'ensemble d'une organisation sur une seule plate-forme et en les modélisant afin d'obtenir une représentation de la réalité facile à interpréter pour les décideurs.

Les données peuvent être issues d'un large éventail d'événements ou d'activités :

1. **Saisie manuelle de données,** comme les formulaires d'enquête collectés et traités par un bureau.
2. **Entrées de capteur,** telles que les scans à une ligne de caisses.
3. **Documents, contenus et médias numériques,** tels que les messages sur les réseaux sociaux.
4. **Activités numériques** enregistrées par des déclencheurs logiciels, comme les clics sur un site Web ou une application.

Avant de pouvoir utiliser les données pour identifier des tendances et des mécanismes de causalité, les données doivent se trouver en un seul endroit et être organisées de manière à pouvoir les gérer.

Le processus de Data Integration

Afin de préparer les données brutes pour l'analyse, plusieurs étapes sont nécessaires :

1. Les données sont collectées depuis des flux de capteurs, des saisies manuelles ou des composants logiciels, puis stockées dans des fichiers ou bases de données.
2. Les données sont extraites des fichiers, bases de données et points de terminaison d'API, puis centralisées dans une destination.
3. Les données sont nettoyées et modélisées pour répondre aux besoins en Analytics des diverses entités de l'entreprise.
4. Les données sont utilisées pour alimenter des produits ou la Business Intelligence.

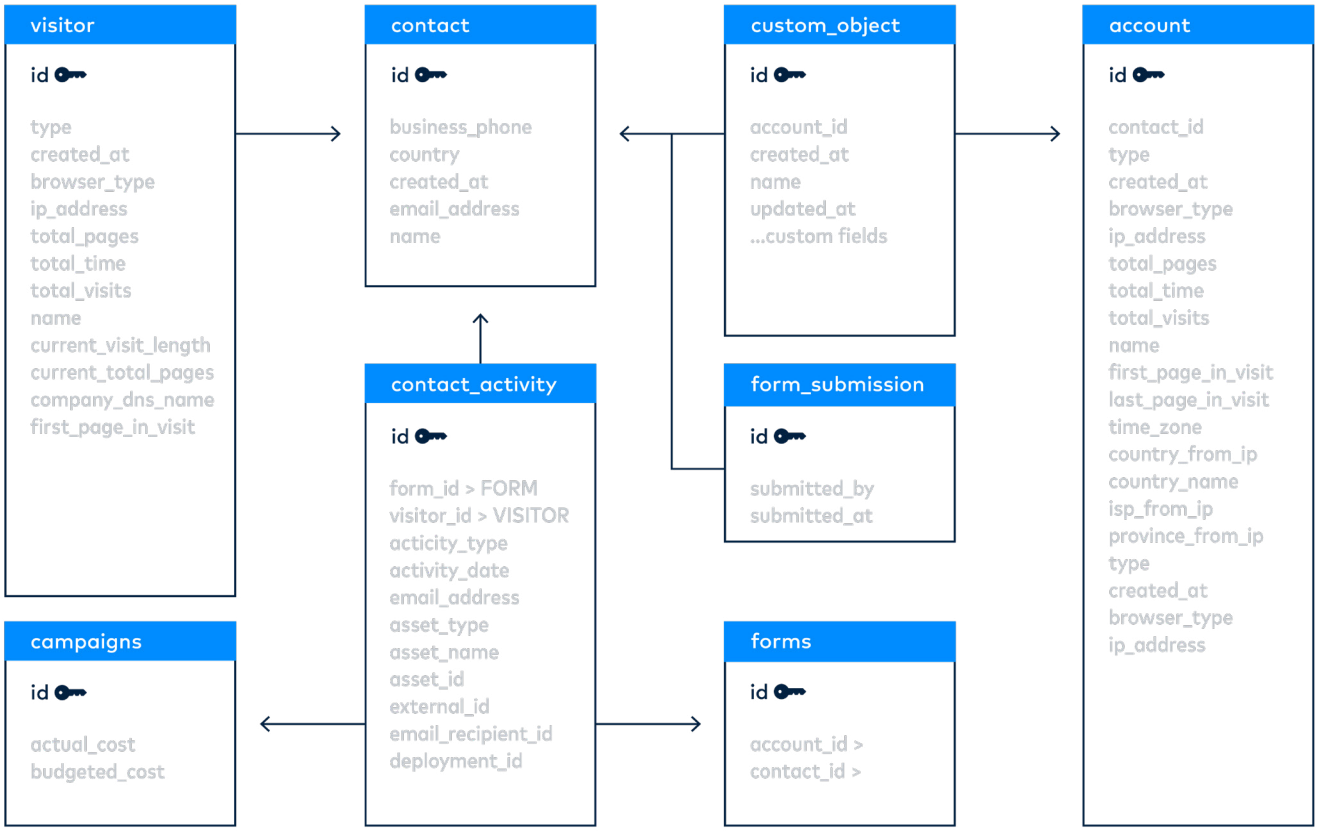
La Data Integration peut être réalisée de façon manuelle, ad hoc ou automatisée à l'aide d'un logiciel. L'approche ad hoc est lente, sujette aux erreurs, non reproductible et non évolutive. Elle est également coûteuse et exige l'attention de professionnels des données compétents et bien rémunérés. Pourtant, **62 % des organisations** utilisent encore des feuilles de calcul pour combiner manuellement et visualiser les données. De nombreux data scientists utilisent cette méthode pour produire des rapports ad hoc.

De son côté, un état d'esprit moderne à l'égard des données ouvre la porte à de bien meilleures façons de travailler. L'approche automatisée de la Data Integration implique une série d'outils appelée Modern Data Stack. Le logiciel de Data Integration utilisé dans une Modern Data Stack peut être développé en interne par l'équipe d'ingénieurs de l'entreprise ou entièrement externalisé et automatisé. Cette approche systématique est plus rapide, plus fiable, plus précise et plus rentable que l'intégration manuelle. Elle apporte également plus de satisfaction aux data engineers, analystes et data scientists qui peuvent alors se consacrer à des tâches gratifiantes plutôt qu'au processus lent et répétitif de collecte et de nettoyage des données.

Caractéristiques importantes des sources et des destinations de données

Les sources de données comprennent les fichiers, les bases de données et les points de terminaison d'API. Dans de nombreux cas, ces éléments sont le produit d'applications. Chaque source de données comporte un modèle de données sous-jacent qui reflète une certaine version de la réalité. Les modèles de données pour les applications, par exemple, montrent comment les utilisateurs interagissent avec le produit. Un **modèle de données** est une représentation abstraite, strictement formatée, de la réalité. Un **schéma** est un plan pratique pour transformer un modèle de données en une base de données avec des tables, des lignes, des colonnes et des interrelations.

Schéma standardisé



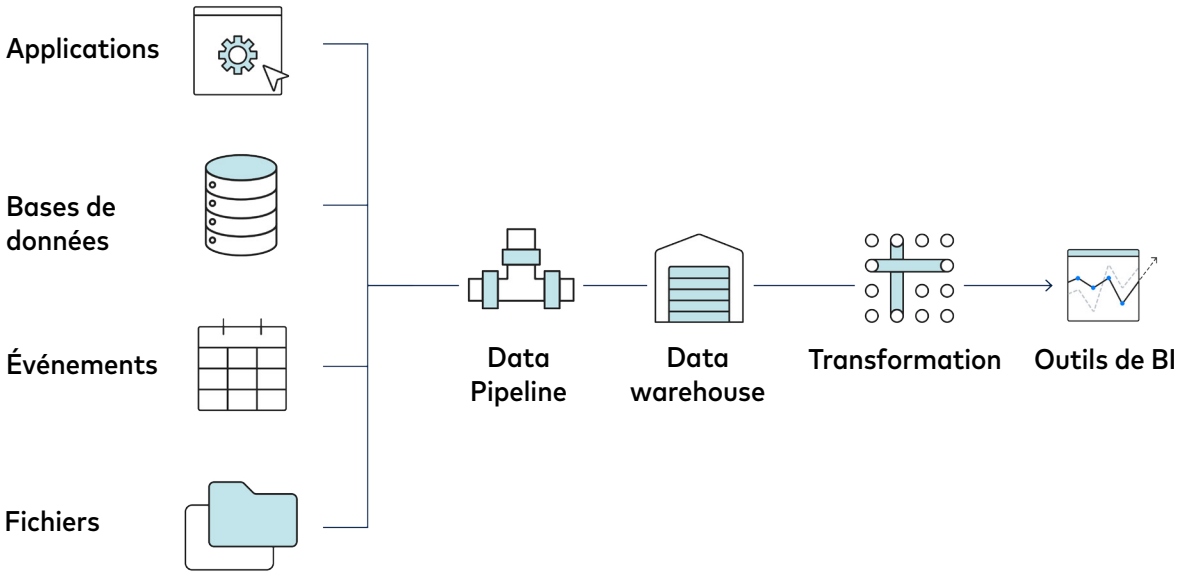
La destination dans laquelle les données sont centralisées pour devenir une source de vérité est généralement une base de données relationnelle optimisée pour l'analyse, appelée **data warehouse** ou entrepôt de données. Les data warehouses stockent des données structurées qui suivent des règles de formatage très spécifiques afin de faciliter leur interprétation par la machine et qui sont organisées en tableaux de lignes et de colonnes.

Dans certains cas, la destination peut être un **Data Lake** ou lac de données. Nous n'aborderons pas ici le sujet des arbitrages entre Data Lakes et data warehouses. Disons pour résumer que les lacs de données sont davantage destinés aux cas d'utilisation qui nécessitent des données non structurées et le stockage en masse de fichiers multimédias ou de documents, comme certaines formes de machine learning. Les technologies plus récentes combinent les fonctionnalités des data warehouses et des Data Lakes dans une plate-forme de données Cloud commune. Cette solution intégrée prend en charge l'Analytics, le machine learning, la réplication des données, l'activation des données et tous les autres besoins liés.

Qu'est-ce qu'une Modern Data Stack ?

Nous avons déjà dit quelques mots sur la Modern Data Stack. Entrons maintenant dans les détails. Une Data Stack est une série d'outils et de technologies qui permet d'automatiser la Data Integration. La Modern Data Stack exploite les dernières évolutions technologiques du Cloud et de l'automatisation. Ses éléments de base sont les suivants :

1. **Sources de données :** elles prennent généralement les formes suivantes :
    a. Applications
    b. Bases de données
    c. Fichiers
    d. Événements numériques
2. **Data pipeline** : logiciel comportant des connecteurs de données qui extraient les données d'une source et les chargent dans une destination. Les connecteurs de données peuvent également appliquer des transformations légères comme la normalisation et le nettoyage des données, orchestrer des transformations dans des modèles pour les analystes ou simplement charger des données brutes.
3. **Destinations** : dépôts centraux de données, généralement des **Data Lakes ou data warehouses**, qui stockent en permanence de grandes quantités de données.
4. **Couche de transformation et de modélisation :** il est généralement nécessaire de transformer les données brutes pour les préparer à l'analyse. Il peut s'agir de **joindre des tables, d'effectuer des calculs d'agrégation, de faire pivoter les données ou de les reformater.** Les transformations peuvent être réalisées dans des environnements de préproduction au sein du Data Pipeline (ETL) ou du data warehouse (ELT).
5. **Outils d'analyse :** ce sont notamment des plates-formes de Business Intelligence prêtes à l'emploi pour le reporting et les tableaux de bord, ainsi que des packages d'Analytics et de science des données pour les langages de programmation courants. Les outils d'analyse sont utilisés pour produire des visualisations, des résumés, des rapports et des tableaux de bord.



On peut voir la Data Integration comme un type particulier d'une opération plus générale appelée mouvement des données. Outre la Data Integration, le mouvement des données comprend également la réplication des données entre les systèmes opérationnels pour des raisons de redondance et de performance. L'**activation des données** transfère les données modélisées à des fins d'analyse d'une destination vers des systèmes opérationnels, ce qui permet toutes sortes d'automatisation des processus métier axés sur les données.

Chapitre 3 : Approches de la Data Integration

Il existe deux approches principales de la Data Integration. L'une de ces architectures, ETL (Extract, Transform, Load ou extraction, transformation, chargement), est très répandue, mais elle est coûteuse et devient rapidement obsolète. L'autre, ELT (Extract, Load, Transform ou extraction, chargement, transformation), est beaucoup plus accessible et tire parti des avancées technologiques régulières.

Avant d'examiner les différences entre les approches, penchons-nous sur les actions qu'elles impliquent.

La transformation expliquée

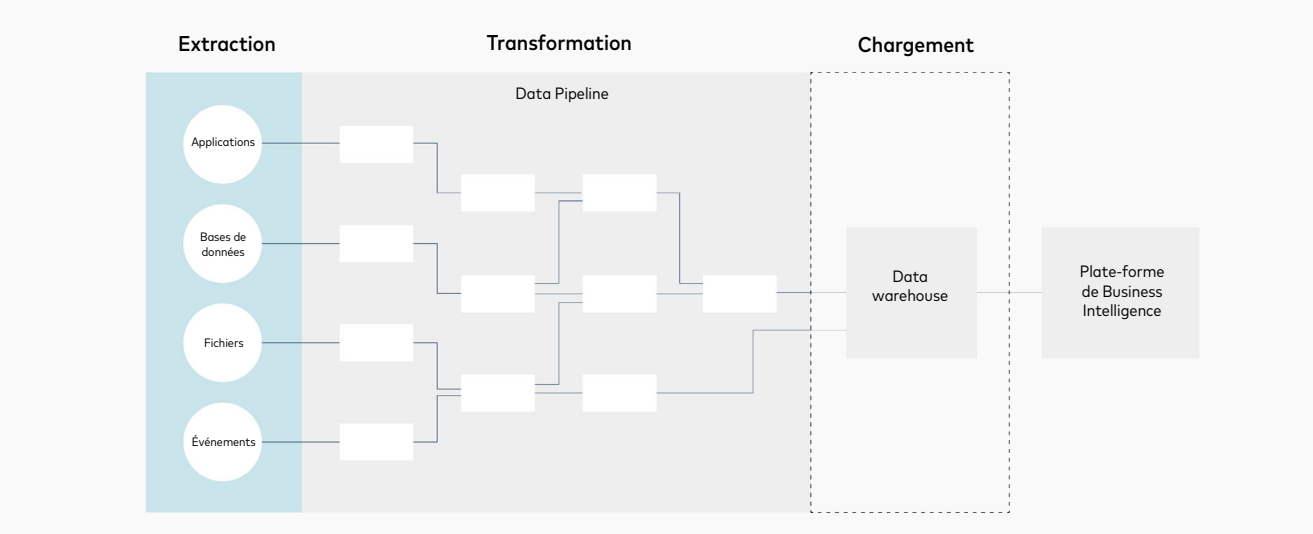
La transformation joue un rôle central dans l'**ETL comme dans l'ELT**. La transformation peut être définie comme l'une des opérations permettant de convertir des données brutes en modèles de données prêts à être analysés. Ces opérations sont les suivantes :

1. **La révision** des données permet de s'assurer que les valeurs sont correctes et organisées pour permettre leur utilisation prévue. Il s'agit par exemple de corriger des fautes d'orthographe, de changer de formats, de hacher des clés, de dédupliquer des enregistrements, entre autres corrections.
2. **Le calcul** implique de calculer des taux, des proportions, des statistiques sommaires et d'autres chiffres importants, ainsi que de transformer des données non structurées en données structurées pouvant être interprétées par un algorithme.
3. **La séparation** consiste à diviser les valeurs en leurs parties constitutives. Les valeurs des données sont souvent combinées dans le même champ en raison d'idiosyncrasies dans la collecte des données, mais il peut être nécessaire de les séparer pour effectuer une analyse plus précise. Par exemple, on peut avoir un champ « adresse » dans lequel le numéro et le nom de la voie, la ville et l'État doivent tous être séparés.
4. **La combinaison** d'enregistrements provenant de différentes tables et sources est essentielle pour obtenir une image complète des activités d'une organisation.

La transformation peut être compliquée et nécessiter des calculs intensifs, en fonction de l'enchaînement et de l'orchestration minutieux des différentes opérations. Elle est très sensible aux changements de schéma, qu'il s'agisse d'une évolution des sources de données en amont ou des besoins commerciaux en aval. Les incidences sur l'évolutivité et l'adaptabilité d'une architecture de Data Integration sont décisives. Le moment où intervient la transformation dans le flux de travail de Data Integration peut faire une énorme différence en termes de temps, de talents et d'argent impliqués dans la construction et la maintenance des Data Pipelines.

Qu'est-ce que l'ETL ?

Datant des années 1970, l'approche traditionnelle de la Data Integration appelée ETL (Extract, Transform, Load ou extraction, transformation, chargement) est tellement répandue que le terme « ETL » est souvent utilisé comme synonyme de Data Integration. Avec la méthode ETL, les Data Pipelines extraient les données des sources, les transforment en modèles de données pour produire des rapports et des tableaux de bord, puis chargent les données dans un data warehouse.

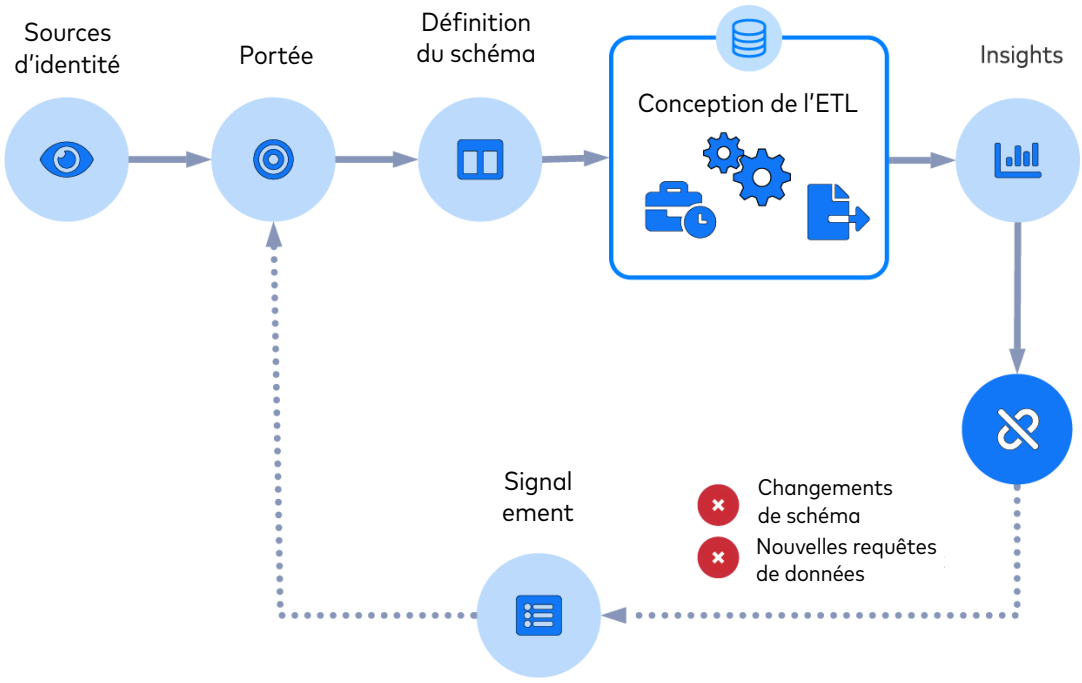


Les processus ETL de transformation des données agrègent ou résument les données, réduisant ainsi leur volume total. En transformant les données avant de les charger, l'ETL limite le volume de données dans le data warehouse, ce qui préserve les ressources de stockage, de calcul et de bande passante. Lorsque l'ETL a été inventé dans les années 1970, les organisations évoluaient dans un contexte de pénurie extrême de stockage, de calcul et de bande passante. Contraintes par la technologie de l'époque, elles n'avaient d'autre choix que de consacrer un temps précieux d'ingénierie à la construction de systèmes ETL sur mesure pour déplacer les données de la source vers la destination.

Le flux de travail projet pour l'ETL se compose des étapes suivantes :

1. Identification des sources de données désirées.
2. Détermination des besoins en Analytics exacts que le projet doit couvrir.
3. Définition du schéma/modèle de données dont les analystes et autres utilisateurs finaux ont besoin.
4. Conception du Pipeline contenant les fonctions d'extraction, de transformation et de chargement. Cela nécessite un investissement important en temps d'ingénierie.
5. Analyse des données pour en extraire des insights.

Flux de travail ETL



Avec l'ETL, les opérations d'extraction et de transformation sont étroitement liées car elles sont effectuées avant que les données ne soient chargées dans une destination. De plus, comme les transformations sont dictées par les besoins spécifiques en Analytics, chaque Pipeline ETL est une solution complexe sur mesure. La personnalisation par nature de ces Pipelines ETL rend l'ajout ou la révision des sources et des modèles de données particulièrement complexe.

Vous commencez probablement à voir les failles et les pièges potentiels de cette approche. Elle était justifiée lorsque les ressources informatiques et de stockage étaient limitées, et acceptable lorsque les sources de données restaient constantes et prévisibles.

Mais le monde a changé. Le flux de travail ETL, y compris tous les efforts déployés pour le définir, le construire et le tester, doit être répété chaque fois que les schémas ou les sources de données en amont changent (notamment lorsque des champs sont ajoutés, supprimés ou modifiés à la source) ou lorsque les besoins en Analytics en aval évoluent, nécessitant de nouveaux modèles de données.

Imaginons qu'une application réorganise les tables de son modèle de données afin de prendre en charge de nouvelles données clients. Le code du Pipeline pour l'extraction, la transformation et le chargement des données dépend de l'ancien schéma et ne fonctionnera plus. Il faudra donc le réécrire. Cet arrêt interrompra toutes les mises à jour, empêchant l'entreprise de prendre des décisions actualisées et informées.

Dans un autre scénario, un analyste peut vouloir créer un nouveau modèle d'attribution qui nécessite de relier plusieurs sources de données de manière inédite. Comme dans le cas précédent, cela se traduit par un processus de reconstruction du flux de travail lent et laborieux.

Étant donné que l'extraction et la transformation précèdent le chargement, tous les arrêts de transformation empêchent le chargement des données vers la destination, ce qui entraîne un temps d'arrêt du Data Pipeline.

Par conséquent, l'utilisation de l'ETL pour la Data Integration présente les difficultés suivantes :

- **Maintenance et révision constantes :** comme le Data Pipeline extrait et transforme les données, dès que les schémas en amont ou les modèles de données en aval changent, le Pipeline s'interrompt et la base de code doit être révisée.
- **Personnalisation et complexité :** les Data Pipelines effectuent des transformations sophistiquées adaptées aux besoins en Analytics spécifiques des utilisateurs finaux. Autrement dit, le code est personnalisé.
- **Perte de ressources d'ingénierie :** la construction et la maintenance du système requièrent des ingénieurs à temps plein du fait que tout fonctionne sur une base de code sur mesure.

Ces difficultés sont encore aggravées si la configuration ETL réside sur site, hébergée dans des centres de données et des fermes de serveurs directement gérées par l'organisation. Ces solutions nécessitent encore plus de réglages et de configuration matérielle.

L'ETL prend le parti de préserver les ressources de calcul et de stockage au détriment du capital humain. Cela avait du sens lorsque les besoins en données étaient beaucoup plus simples et que les ressources de calcul et de stockage étaient extrêmement rares, en particulier par rapport à la main-d'œuvre. Les circonstances ont changé, et l'ETL est aujourd'hui une option extrêmement coûteuse et à forte intensité de main-d'œuvre dont les avantages sont discutables.

Les tendances technologiques changent la donne et rendent l'ETL obsolète

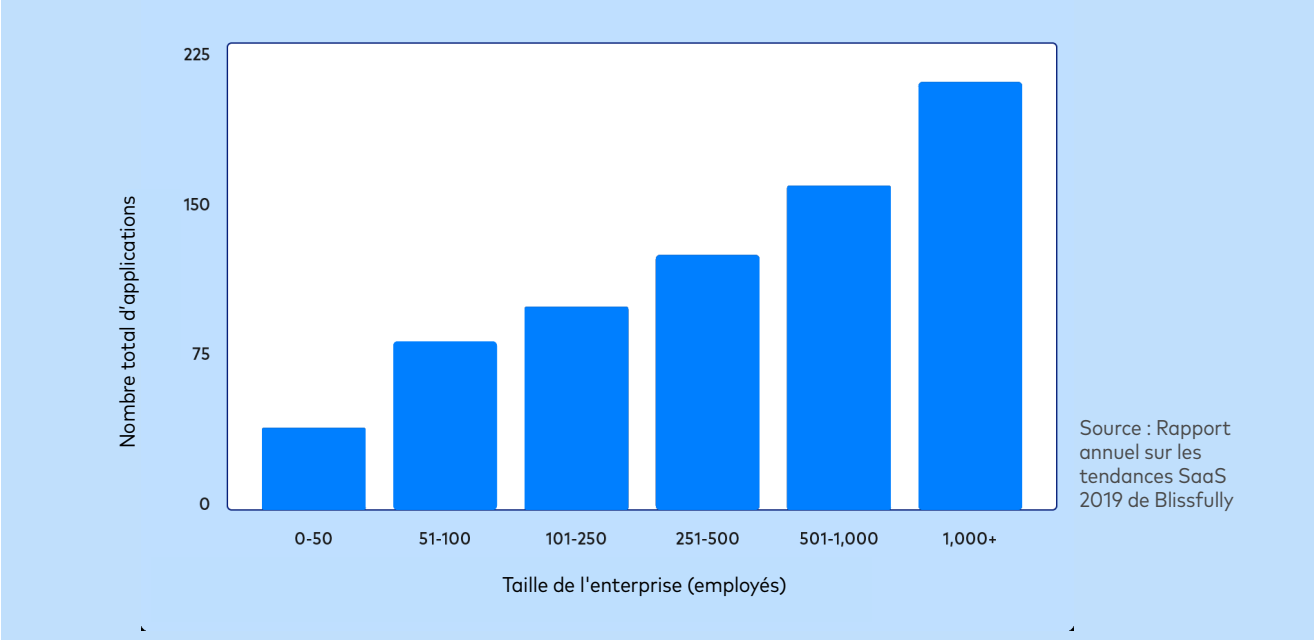
Ces dernières années, le « big data » et le « Cloud » sont devenus des mots à la mode, et ils sont tous deux directement liés à la disparition de l'ETL en tant que méthode pratique d'intégration et d'analyse des données. Le big data fait référence au volume massif et à la complexité des données modernes, et résulte de la croissance du calcul et du stockage décentralisés sur Internet sous la dénomination de « Cloud ». Les applications et les appareils basés sur le Cloud (y compris l'**Internet des objets** en pleine expansion) produisent de vastes empreintes numériques de données qui peuvent être transformées en insights de valeur.

Ces données sont généralement stockées dans des fichiers et des bases de données opérationnelles basés sur le Cloud, puis exposées aux utilisateurs finaux sous différentes formes :

1. Flux d'API
2. Systèmes de fichiers
3. Journaux de bases de données et résultats de requêtes
4. Flux d'événements

Les technologies basées sur le Cloud, en particulier le Software-as-a-Service (SaaS), sont de plus en plus répandues. Une organisation type utilise désormais des dizaines voire des centaines d'applications.

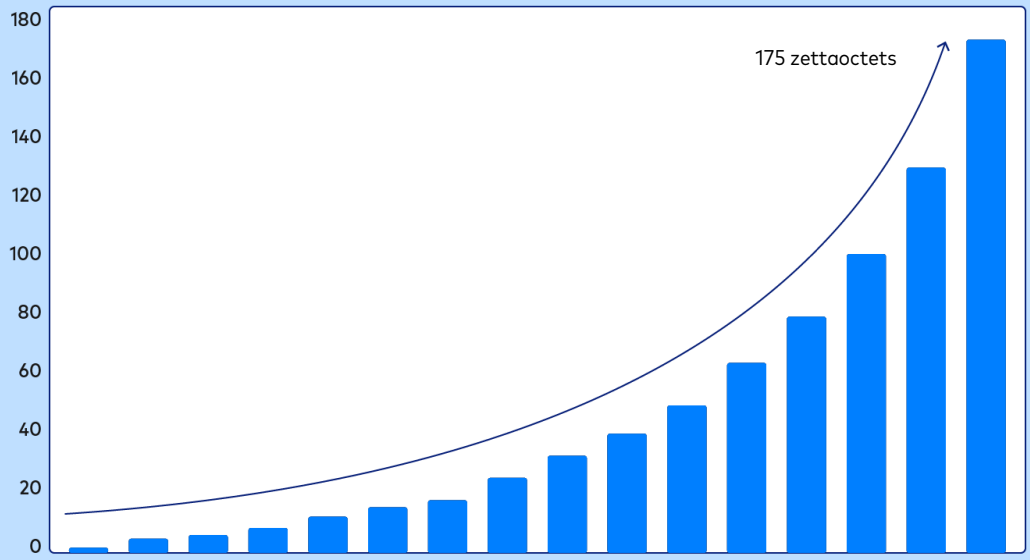
Nombre d'applications par entreprise



Source : Rapport annuel sur les tendances SaaS 2019 de Blissfully

En conséquence, le volume global de données produites dans le monde a explosé au cours des dernières décennies :

Taille annuelle de la sphère de données mondiale

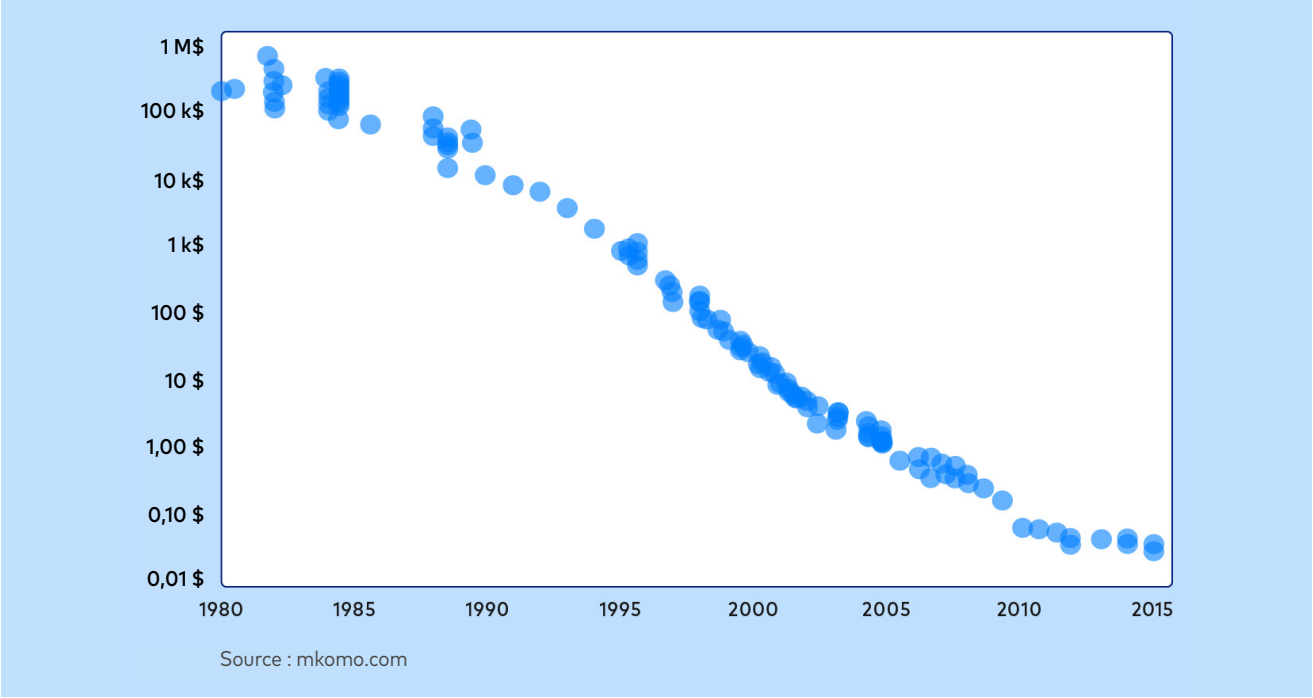


Source : « Data Age 2025 », commandité par Seagate sur la base de l'étude IDC Global DataSphere, novembre 2018

Le volume et la précision des données ne cessent de croître, tout comme les possibilités d'analyse. La multiplication des données offre la possibilité de comprendre et de prévoir les phénomènes avec plus de précision que jamais.

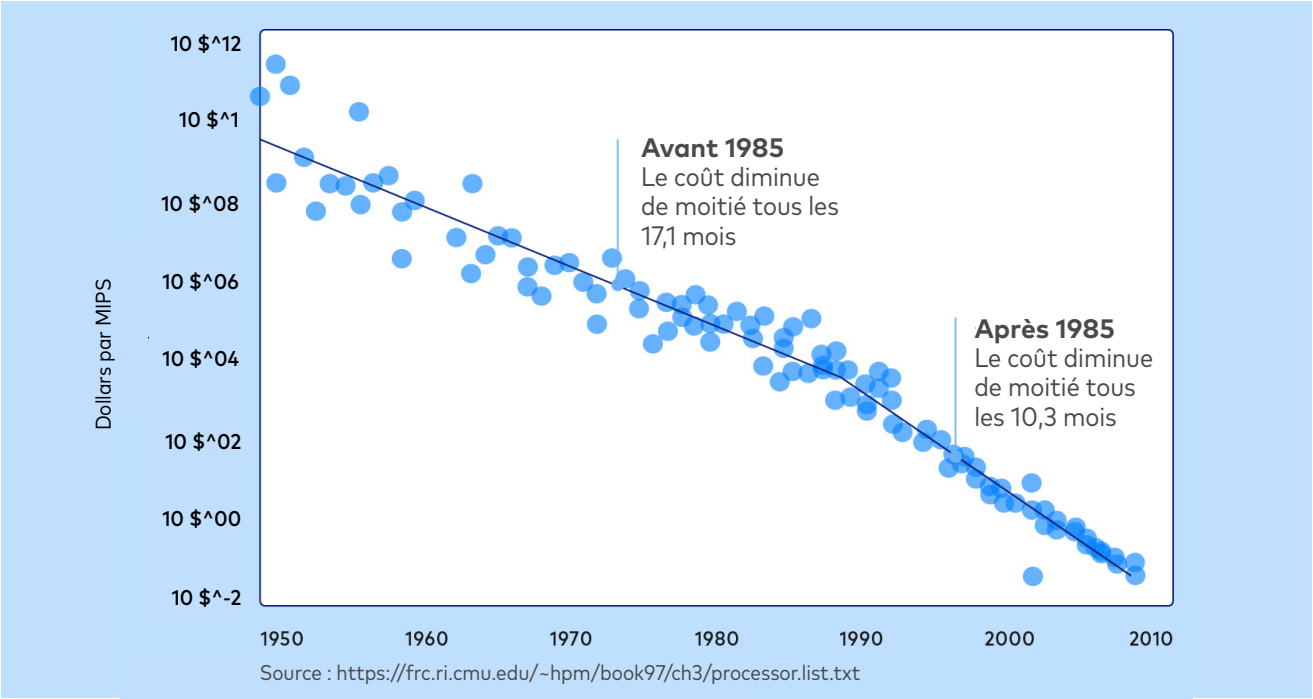
L'ETL avait du sens à une époque où les ressources de calcul, de stockage et de bande passante étaient extrêmement rares et coûteuses. Ces contraintes technologiques ont disparu depuis. Le coût du stockage a chuté de près d'un million de dollars par gigaoctet à quelques centimes en quatre décennies :

Coût des disques durs par gigaoctet



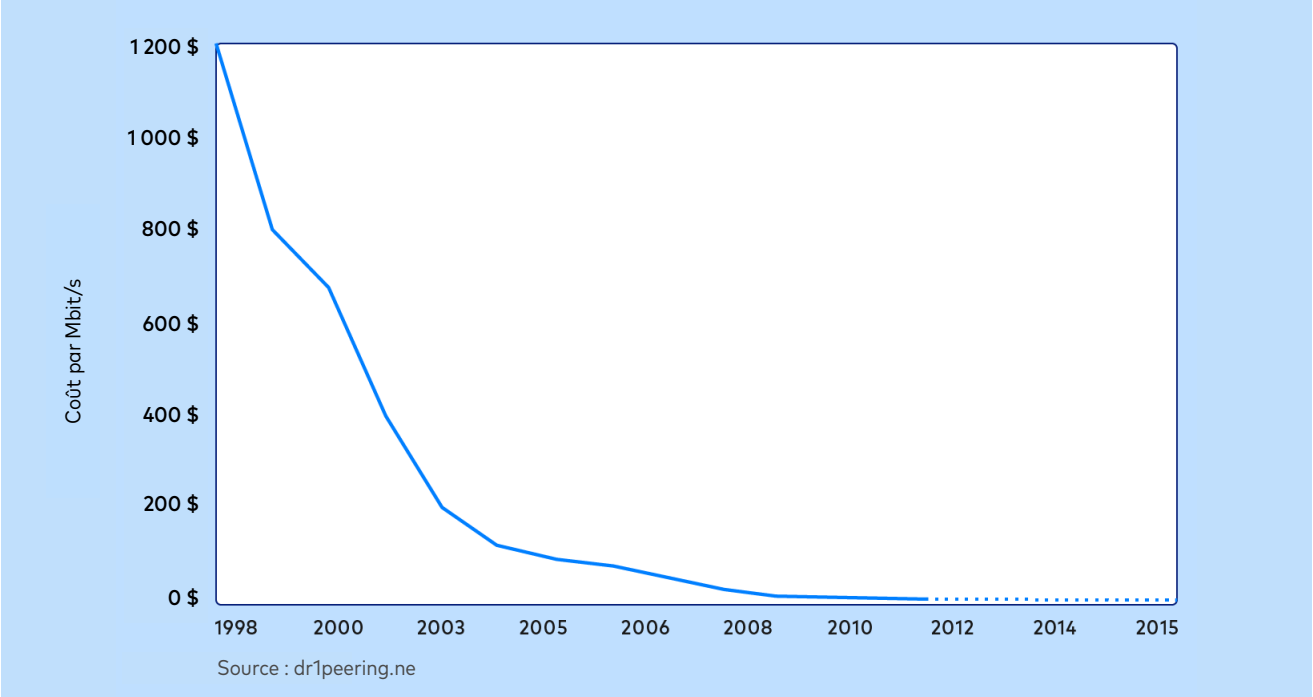
Le coût du calcul a été divisé par plusieurs millions depuis les années 1970 :

Coût du calcul



Le coût du transit Internet a quant à lui été divisé par un facteur de l'ordre des milliers :

Prix du trafic internet (bande passante)



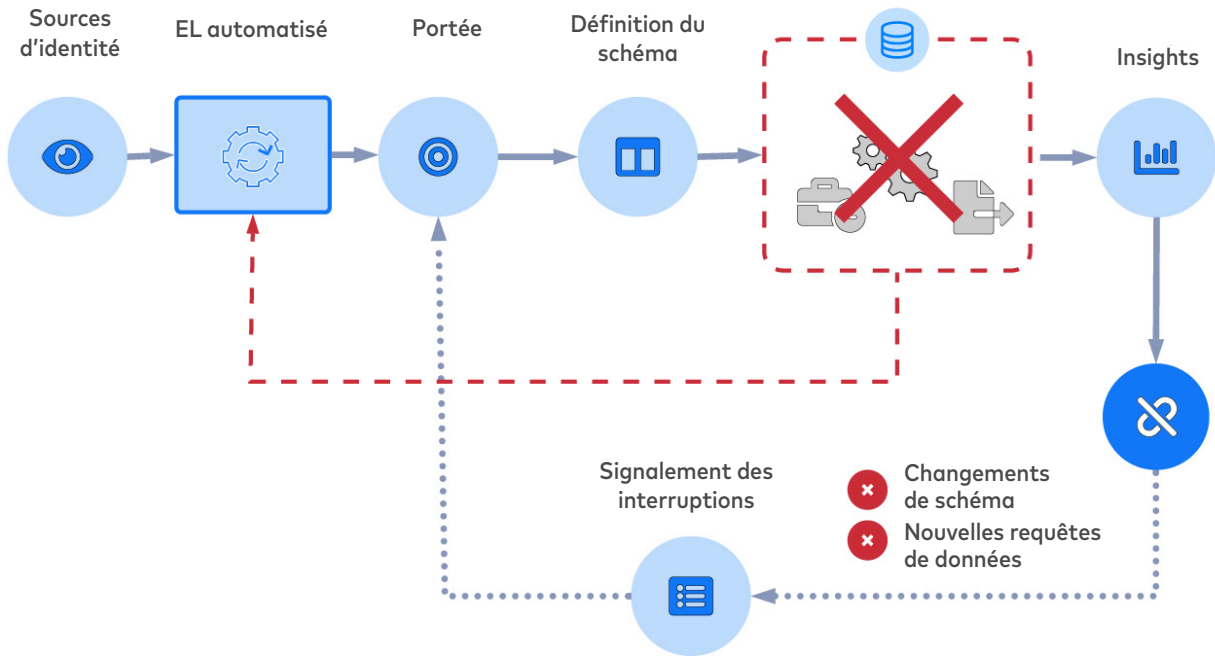
Ces tendances ont rendu obsolètes l'ETL et les efforts coûteux et gourmands en main-d'œuvre qu'il nécessite pour deux raisons. Premièrement, l'accessibilité du calcul, du stockage et de la bande passante Internet a engendré une croissance massive du Cloud et des services basés sur le Cloud. À mesure que le Cloud s'est développé, le volume, la variété et la complexité des données ont crû également. Un Pipeline fragile qui intègre un volume de données et un niveau de précision limités ne suffit plus dans ce contexte.

Deuxièmement, le coût abordable du calcul, du stockage et de la bande passante Internet permet d'héberger les technologies modernes de Data Integration dans le Cloud et de stocker de grands volumes de données non transformées dans des data warehouses. Il est ainsi possible de réorganiser le flux de travail de Data Integration et de réaliser des économies d'argent et de personnel considérables.

ELT : une alternative moderne à l'ETL

La possibilité de stocker d'énormes quantités de données non transformées dans les data warehouses rend possible une nouvelle architecture de Data Integration, l'ELT (Extract, Load, Transform ou extraction, chargement, transformation), dans laquelle l'étape de transformation n'intervient qu'à la fin du flux de travail et les données sont chargées dans une destination plus ou moins immédiatement après l'extraction.

Flux de travail ETL

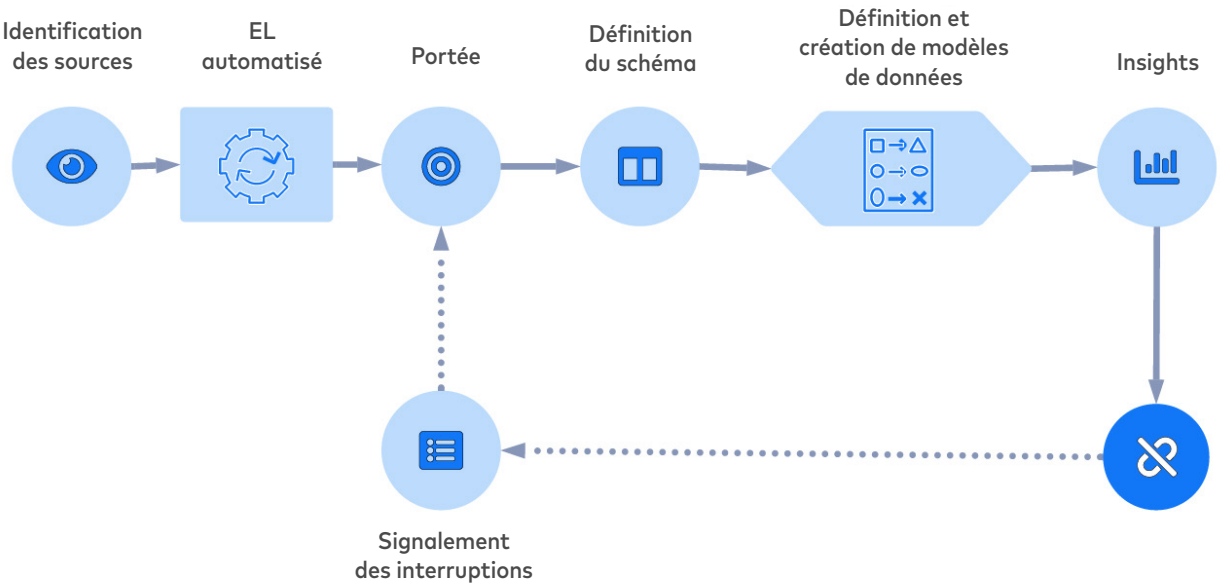


Les deux états de défaillance de l'ETL (à savoir les changements de schéma en amont et de modèle de données en aval) ne peuvent plus interférer avec l'extraction et le chargement, ce qui favorise une approche plus simple et plus fiable de la Data Integration.

Le flux de travail ELT présente un cycle de projet plus court que celui de l'ETL :

1. Identification des sources de données désirées
2. Extraction et chargement automatisés
3. Détermination des besoins en Analytics exacts que le projet doit couvrir
4. Création de modèles de données par conception des transformations
5. Opérations analytiques et extraction d'insights

Flux de travail ELT

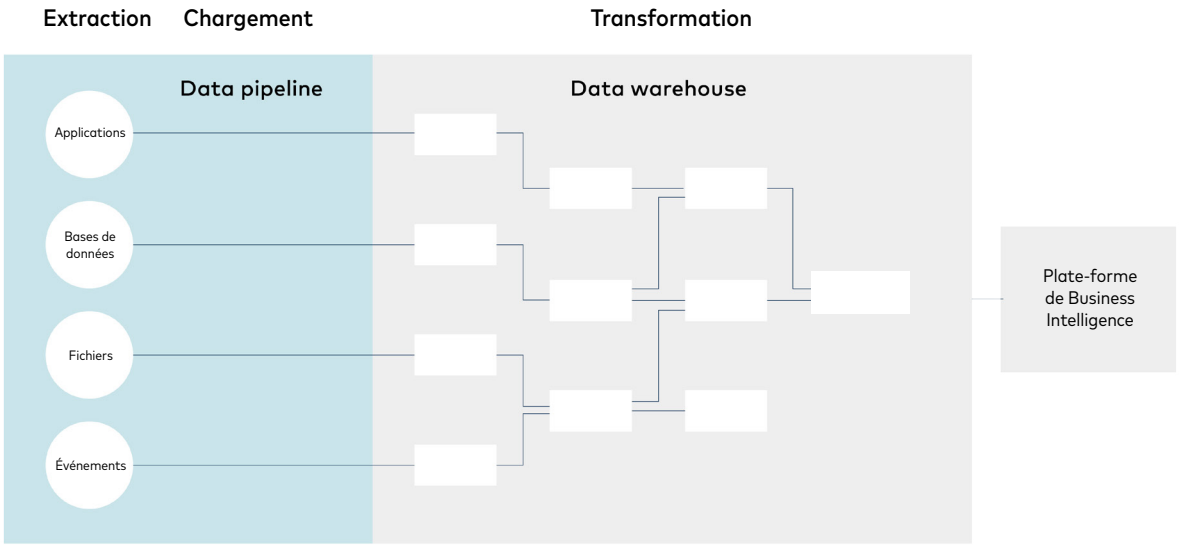


Avec l'ELT, l'extraction et le chargement des données sont indépendants de la transformation car ils interviennent en amont. Même si la couche de transformation peut échouer en cas de changements de schéma en amont ou de modèle de données en aval, ces défaillances n'empêchent pas le chargement des données dans une destination. Une organisation peut continuer à extraire et à charger des données même si les transformations sont périodiquement réécrites par les analystes. Comme les données sont stockées avec peu d'altération, elles constituent une source de vérité exhaustive et actualisée.

En outre, les transformations étant réalisées au sein de l'environnement du data warehouse, il n'est plus nécessaire de les concevoir via des interfaces de type glisser-déposer, de les écrire sous forme de scripts Python ou de créer des orchestrations complexes entre des sources de données disparates. À la place, les transformations peuvent être écrites en SQL, le langage natif de la plupart des analystes. La Data Integration passe ainsi d'une activité relevant du service IT ou de l'ingénierie, caractérisée par de longs cycles de projet et une forte implication des ingénieurs, à une activité en libre-service directement prise en charge par les analystes.

Un autre avantage de l'exécution des transformations dans l'environnement du data warehouse est l'amélioration des performances. Les data warehouses basés sur le Cloud peuvent dimensionner des capacités de calcul et de stockage supplémentaires en fonction des besoins, ce qui facilite et accélère les transformations par rapport aux couches d'environnements de préproduction utilisées dans l'ETL. Le Cloud permet le provisionnement de ressources en mode « juste à temps », ce qui évite les dépenses liées au maintien d'une capacité matérielle excédentaire qui n'est utilisée qu'occasionnellement.

Enfin et surtout, l'architecture ELT simplifie l'externalisation, l'automatisation et la normalisation des opérations d'extraction et de chargement. Comme les transformations s'effectuent au sein du Warehouse, la partie « EL » du Pipeline n'a pas besoin de produire une sortie différenciée en fonction des besoins spécifiques de l'organisation.



Un fournisseur externe tel que Fivetran peut fournir un schéma standardisé à chaque client. Les **moyens de normaliser** un schéma étant assez peu nombreux, la manière la plus pertinente de standardiser un modèle de données est la **normalisation**. La normalisation favorise l'intégrité des données, garantissant qu'elles sont exactes, complètes, cohérentes et de provenance connue. Elle facilite également l'interprétation des modèles de données par les analystes. Les résultats normalisés présentent aussi l'avantage de permettre la création de produits dérivés comme des **modèles de modélisation de données normalisés** pour l'Analytics.

En résumé, l'ELT transforme les modèles de données en produits indifférenciés, ce qui permet l'externalisation et l'automatisation. L'équipe chargée des données peut alors passer de la conception, de la maintenance et de la révision d'un logiciel complexe à la simple utilisation des résultats et à la création de modèles de données directement dans la destination.

Principales différences entre ETL et ELT

Le tableau suivant résume les différences entre l'ETL et l'ELT :

ETL	ELT
Extraction, transformation, chargement	Extraction, chargement, transformation
Intégration de résumés ou de sous-ensembles de données	Intégration de toutes les données brutes
Chargement et transformation étroitement couplés	Chargement et transformation non couplés
Temps plus long pour charger les données	Temps plus court pour charger les données
Les échecs de transformation arrêtent le Pipeline	Les échecs de transformation n'arrêtent pas le Pipeline
Prévision des cas d'utilisation et conception des modèles de données à l'avance ou révision complète du Data Pipeline	Création de nouveaux cas d'utilisation et conception de modèles de données à tout moment
Personnalisé	Commercial
Conception et maintenance en continu	Automatisé
Économise des ressources de calcul et stockage	Économise du personnel
Utilisation de langages de script pour les transformations	Utilisation de SQL pour les transformations
Axé sur l'ingénierie/l'IT ; système expert	Axé sur les analystes ; accessible aux non-experts
Dans le Cloud ou sur site	Presque uniquement dans le Cloud

Dans certains cas, l'ETL doit tout de même être préféré à l'ELT. Par exemple, dans les cas suivants :

1. Les modèles de données souhaités sont bien connus et ne sont pas amenés à changer rapidement. En particulier lorsqu'une organisation conçoit et assure la maintenance des systèmes qu'elle utilise en tant que données sources.
2. Des exigences de conformité réglementaire et de sécurité concernant les données interdisent de stocker celles-ci dans un emplacement tiers, tel qu'un espace de stockage dans le Cloud.

Ces conditions sont caractéristiques des très grandes entreprises, des acteurs qui opèrent dans des secteurs hautement réglementés et des organisations spécialisées dans les produits Software-as-a-Service. Dans de tels cas, il peut être pertinent d'utiliser l'ELT automatisé et externalisé pour intégrer les données provenant de sources tierces tout en procédant à la conception et à la maintenance de l'ETL pour intégrer les sources de données internes propriétaires.

Les avantages de l'ELT et de l'automatisation

Une organisation qui allie l'automatisation à l'ELT peut considérablement simplifier son flux de travail de Data Integration. L'ELT automatisé est un multiplicateur de force pour l'ingénierie des données. Cette technologie permet en effet aux équipes de se concentrer sur des projets plus stratégiques, comme l'optimisation de l'infrastructure de données de l'organisation ou la mise en production de modèles de données, plutôt que sur la construction et la maintenance de Data Pipelines. En effectuant le travail de plusieurs ingénieurs à temps plein, l'ELT automatisé permet d'alléger les équipes et d'utiliser les effectifs de manière beaucoup plus efficace.

La reproductibilité et une qualité constante sont d'autres avantages de l'automatisation par rapport à l'intégration manuelle des données. Les analystes et les data scientists peuvent enfin exploiter leur expertise pour modéliser et analyser les données **au lieu de se contenter de les préparer ou de les convertir d'un format à un autre.**

Chapitre 4 : Construction vs acquisition pour la Data Integration

Choisir entre concevoir sa propre intégration de données ou acquérir une solution en tant que service suppose des arbitrages. Cependant, dans la plupart des cas, l'achat d'une solution prête à l'emploi est plus pertinent.

Coût de construction d'un Data Pipeline

Une enquête réalisée par **Wakefield Research** auprès de grandes entreprises (plus de 2 500 employés) aux États-Unis, au Royaume-Uni, en Allemagne et en France a révélé que la création et la maintenance de Data Pipelines coûtaient en moyenne près de 520 000 $ par an. À un coût annuel moyen de 98 000 $ chacun, cela représente le travail de plus de cinq data engineers à temps plein.

Voici un autre calcul pour un cas d'utilisation plus simple. Vous pouvez suivre la démonstration ci-dessous ou **utiliser notre calculateur.**

Afin d'estimer le coût de la construction de Data Pipelines, nous avons besoin des éléments suivants :

1. Coût annuel moyen de la main-d'œuvre pour vos data engineers, analystes ou data scientists
2. Nombre de sources de données dont vous disposez
3. Temps nécessaire à la création et à la maintenance d'une source de données type

Ces données permettent d'estimer le temps et les dépenses consacrés à l'ingénierie.

- Supposons que le coût soit de 140 000 $ pour chaque data engineer, soit un salaire de base de 100 000 $ multiplié par 1,4 pour les avantages sociaux.

Supposons que la construction d'un connecteur prenne environ 7 semaines plus environ 2 semaines par an pour la mise à jour et la maintenance. Chaque connecteur représente 9 semaines de travail par an.

- Pour sept connecteurs, ce sont 7 x (7 + 2) = 63 semaines de travail par an.

À partir du nombre de semaines de travail par an, calculez quelle proportion d'une année de travail cela représente. Multipliez ensuite le résultat par le coût de la main-d'œuvre pour obtenir le coût monétaire total. Considérons que l'année de travail dure 48 semaines après déduction des jours fériés, des congés payés et autres temps d'inactivité.

- Si le coût de la main-d'œuvre est de 140 000 $, que sept connecteurs nécessitent 63 semaines de travail et qu'une année compte 48 semaines de travail, alors (140 000 $) x (63 / 48) = 183 750 $.

D'après notre expérience chez Fivetran, ces chiffres sont des estimations réalistes pour comprendre le coût d'une solution de Data Integration maison. En moyenne, nous constatons que nos clients économisent l'équivalent d'environ deux salaires d'ingénieurs. Cela représente au minimum un montant proche du seuil des six chiffres, même pour un cas d'utilisation relativement simple impliquant un petit nombre de sources de données.

Plus important, l'utilisation du temps des ingénieurs pour la construction et la maintenance des Data Pipelines entraîne des coûts d'opportunité importants. Les ingénieurs, les analystes et les data scientists peuvent mener de nombreux projets de données à forte valeur en aval des Data Pipelines, par exemple :

1. Modèles de données
2. Visualisations, tableaux de bord et rapports
3. Systèmes de production destinés aux clients
4. Automatisation des processus métier
5. Flux d'API et autres produits de données pour les tiers
6. Data Pipelines personnalisés vers des sources pour lesquelles il n'existe pas de Pipelines standard
7. Infrastructure de soutien à l'activation des données
8. Modèles prédictifs, intelligence artificielle et machine learning, à la fois à usage interne et pour les produits destinés aux clients

Éviter l'iceberg de l'intégration : les risques techniques de la construction d'un Data Pipeline

Concevoir un Data Pipeline, c'est comme regarder un iceberg. À première vue, tout est simple. Il suffit de déplacer les données d'un emplacement vers un autre.

Mais comme dans un iceberg, la majeure partie du travail nécessaire est cachée, tapie sous la surface. La Data Integration est compliquée et exige un système complexe pour résoudre un large éventail de problèmes techniques dans les domaines suivants :

- Automatisation
- Performances
- Fiabilité
- Évolutivité
- Sécurité

Automatisation

L'objectif de la construction d'un Data Pipeline est de déplacer de manière programmée les données d'une source vers une destination au lieu de manipuler manuellement des fichiers. Pour ce faire, vos data engineers doivent concevoir et construire un système facile à paramétrer pour les utilisateurs de façon à extraire, transformer et charger des données à intervalles réguliers, sans configuration ni déclenchement manuels.

La construction entière de ce système implique également d'organiser les données brutes de la source dans une structure que les analystes peuvent utiliser pour produire des tableaux de bord et des rapports. Cela nécessite une compréhension approfondie des données sous-jacentes.

Enfin, vous devez créer des outils permettant de surveiller l'état d'un Pipeline et d'alerter vos équipes dès qu'un problème survient afin de corriger les erreurs et les arrêts, ce qui nous amène aux points suivants.

Performance

Un bon Data Pipeline doit être performant et fournir les données avant qu'elles ne deviennent trop anciennes pour être exploitables. Il doit également éviter d'interférer avec les opérations métier critiques pendant son fonctionnement.

Une approche consiste à capturer les modifications de manière incrémentielle à l'aide des journaux, des horodatages ou des déclencheurs au lieu d'interroger des tables ou des schémas de production entiers. On parle de « capture des changements de données », en particulier lors de la lecture à partir de bases de données.

Parmi les autres considérations liées aux performances, citons l'exécution en parallèle et la distribution de l'architecture du Data Pipeline afin d'utiliser davantage de ressources

selon les besoins, ce qui permet au système de faire face à un pic de charge ou de demande. L'augmentation ou la diminution des capacités de calcul et de stockage et la budgétisation adéquate de ces activités peuvent s'avérer complexes. Il convient également d'identifier et de réduire les goulots d'étranglement par un cloisonnement et la mise en cache des opérations sensibles.

Fiabilité

La performance prend tout son sens lorsque la fiabilité des données est au rendez-vous. La synchronisation des données d'un emplacement vers un autre peut échouer pour de nombreuses raisons. L'une des plus courantes est un simple changement de schéma en amont. Selon la manière dont le Data Pipeline est construit, l'ajout ou la suppression de tables et de colonnes peut le faire complètement dérailler.

Des bugs et des défaillances matérielles de toutes sortes sont également fréquents. Votre Pipeline peut subir des fuites de mémoire, des pannes de réseau, des échecs de requêtes, etc. Afin de récupérer les synchronisations qui ont échoué, vous devez construire un système **idempotent**, dans lequel les opérations peuvent être répétées pour produire le même résultat après l'application initiale.

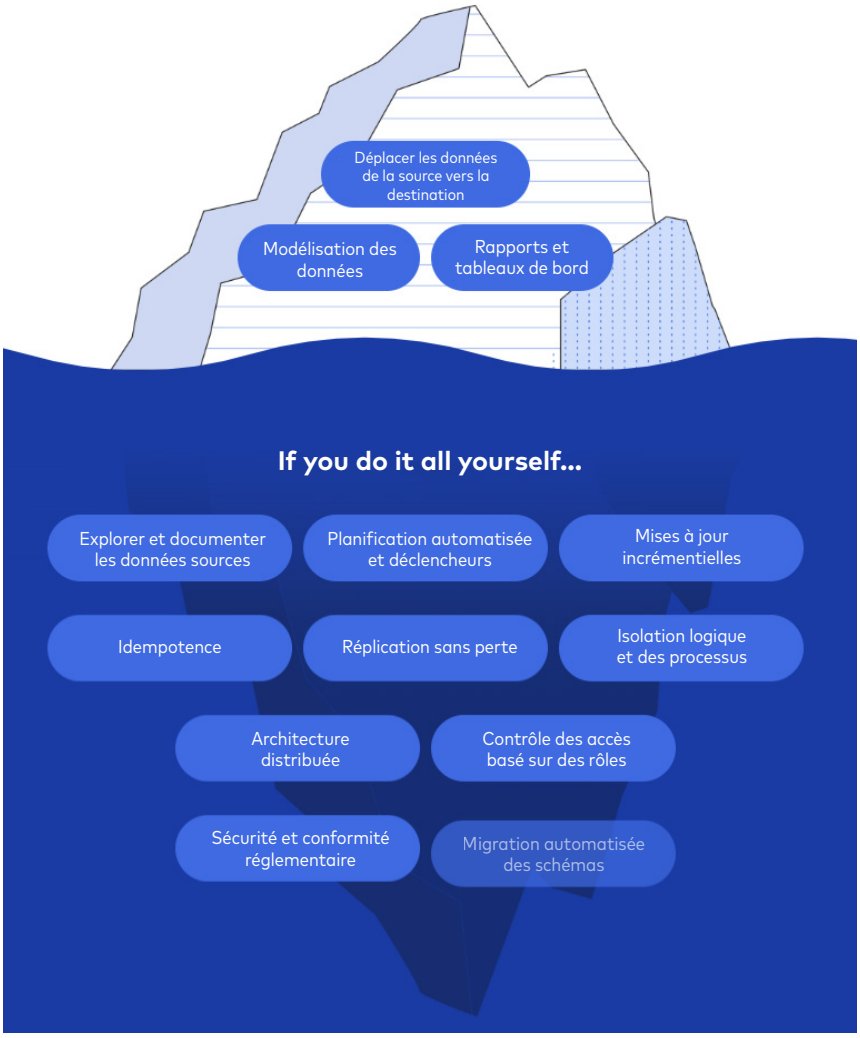
Évolutivité

À mesure que votre organisation se développe, elle devra prendre en charge un nombre croissant de sources de données et un volume de données plus important. Il y a de fortes chances que vous soyez également confronté à des exigences de performances plus élevées, par exemple des délais d'exécution plus courts, à mesure que l'utilisation des données par votre organisation progresse. Votre charge d'ingénierie sera d'autant plus lourde que vous devrez construire et gérer des connecteurs pour chaque nouvelle source de données.

Alors que le nombre de vos connecteurs et utilisateurs augmente, peut-être devrez-vous aussi envisager un système de contrôle automatisé de votre Data Pipeline.

Sécurité

Pour des raisons de conformité réglementaire dans de nombreuses juridictions, votre Data Pipeline ne doit pas exposer ni stocker des informations permettant d'identifier personnellement un utilisateur. L'ensemble du trafic sur votre infrastructure doit être chiffré et votre système doit appliquer l'isolation logique et des processus pour garantir qu'aucune donnée sensible n'est envoyée par erreur à la mauvaise destination.



Coûts prévisibles d'une solution automatisée

L'alternative à la construction d'un Data Pipeline maison est d'en acheter un. L'abonnement à une solution automatisée peut être facturé selon un forfait annuel ou un tarif basé sur la consommation. Il existe de nombreux modes de tarification, par exemple les barèmes basés sur le nombre mensuel de lignes actives (MAR).

L'automatisation vise essentiellement à échanger de l'argent contre une somme beaucoup plus importante en main-d'œuvre et en temps. Une bonne solution peut vous permettre de commencer à synchroniser les données dans le Warehouse en quelques minutes ou quelques heures plutôt qu'en plusieurs semaines ou plusieurs mois. Plus important, elle doit vous permettre de faire beaucoup plus avec moins d'argent et de talents. Une bonne solution prête à l'emploi peut facilement remplacer un ou deux data engineers.

Chapitre 5 : Comment concevoir une Modern Data Stack

Nous avons précédemment défini une Data Stack comme l'ensemble des outils et des processus utilisés pour extraire, charger, transformer et analyser les données. La Modern Data Stack tire parti des progrès des technologies basées sur le Cloud, des outils tiers et de l'automatisation. Comme décrit précédemment, les éléments essentiels d'une Modern Data Stack sont les suivants :

1. Outil de Data Integration
2. Data warehouse
3. Plate-forme de Business Intelligence

Ces éléments peuvent également inclure une prise en charge des transformations, de l'activation des données et du machine learning. Ensemble, les éléments basés sur le Cloud d'une Modern Data Stack simplifient radicalement la mise en œuvre des décisions fondées sur les données par rapport aux technologies traditionnelles sur site ou anciennes.

Considérations commerciales clés

Pour chaque élément de votre Data Stack, tenez compte des facteurs suivants :

1. **Prix et coûts :** assurez-vous que les barèmes de prix et de coûts de chaque outil sont pertinents pour votre organisation. Soyez attentif au coût total de possession ainsi qu'aux coûts d'opportunité des alternatives, notamment des solutions maison.
2. **Adaptation aux compétences et aux projets de l'équipe :** votre équipe doit être capable d'utiliser les outils et les technologies en question. Par exemple, vos analystes sont peut-être mieux armés pour effectuer des transformations en utilisant SQL plutôt qu'un langage de script.
3. **Dépendance vis-à-vis d'un fournisseur et anticipation des besoins :** pourrez-vous continuer à intégrer les données si un fournisseur cesse son activité ou modifie ses conditions de service ?

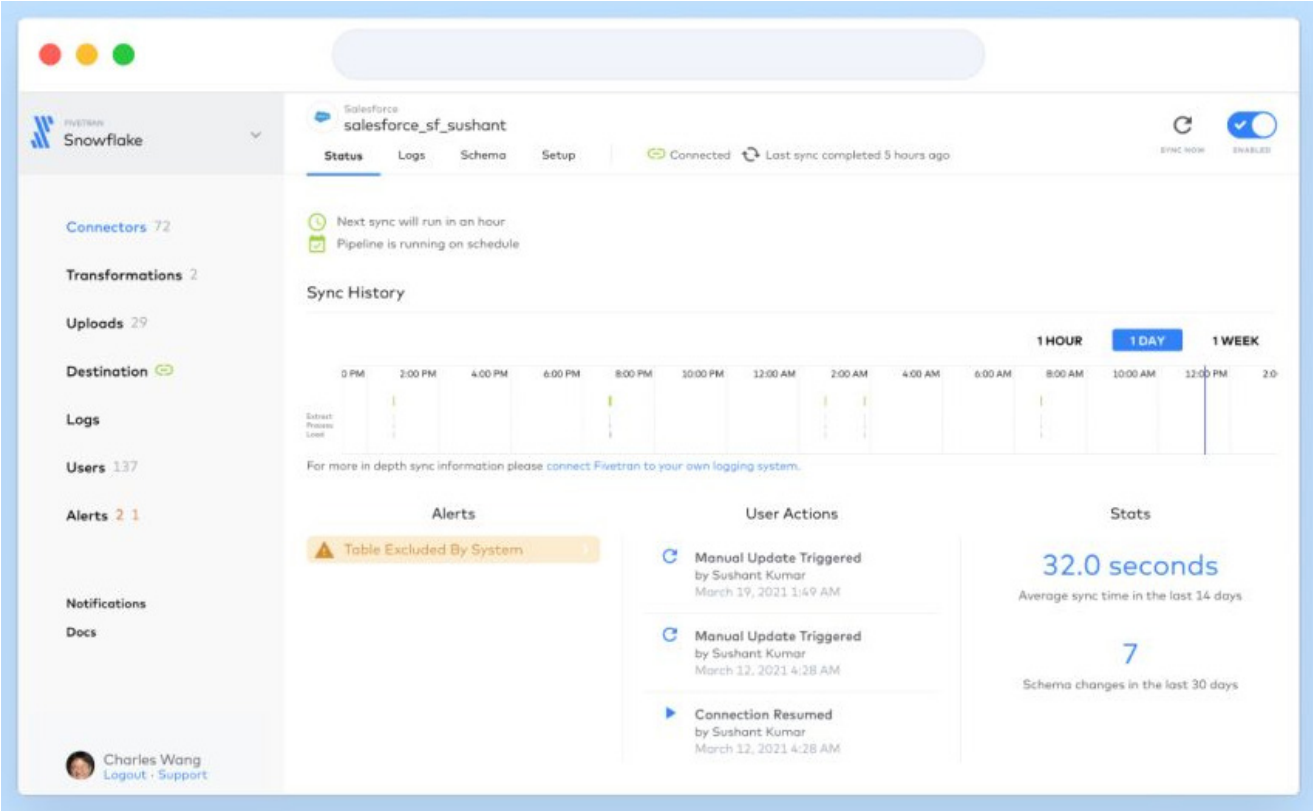
Ces considérations organisationnelles sont davantage liées à la manière dont vous envisagez de développer et de soutenir votre organisation à l'avenir qu'aux caractéristiques techniques de chaque outil.

Choisir le bon outil de Data Integration

Il existe de nombreux outils de Data Integration sur le marché, et leurs approches techniques comme leurs fonctionnalités varient considérablement. Voici les principaux facteurs à prendre en compte lors du choix d'un outil de Data Integration :

1. **Qualité des connecteurs de données :** tenez compte de ces facteurs lorsque vous évaluez la qualité des connecteurs :
    a. **Open source vs propriétaire :** il existe des connecteurs open source pour une plus large variété de sources de données, mais leurs équivalents propriétaires sont souvent de meilleure qualité et s'intègrent plus facilement aux autres composants d'une Data Stack.
    b. **Schémas standardisés et normalisation :** les données provenant des flux d'API ne sont généralement pas normalisées. La normalisation soutient l'intégrité des données, tandis que la standardisation favorise l'externalisation et l'automatisation en permettant à un fournisseur de proposer la même solution à un large éventail de clients.
    c. **Mises à jour incrémentielles vs complètes :** les mises à jour incrémentielles à partir des journaux ou d'autres formes de détection des changements permettent des mises à jour plus fréquentes qui n'interfèrent pas avec les opérations métier.
2. **Compatibilité avec les sources et destinations :** l'outil est-il compatible avec vos sources et vos destinations ? Le fournisseur offre-t-il un moyen pour les clients de suggérer de nouvelles sources et destinations ? En ajoute-t-il régulièrement de nouvelles ?
3. **Configuration vs zero-touch :** les outils entièrement gérés ne nécessitant aucune intervention sont extrêmement accessibles, avec des connecteurs standardisés, testés sous contrainte et sans maintenance. À l'inverse, les outils configurables nécessitent une allocation coûteuse de temps d'ingénierie.
4. **Automatisation :** les outils d'intégration doivent supprimer autant d'interventions et d'efforts manuels que possible. Vérifiez si un outil offre des fonctionnalités telles que la migration automatisée des schémas, l'adaptation automatique aux modifications d'API et la programmation continue de la synchronisation. Les machines coûtent généralement moins cher que les êtres humains, et le but de l'automatisation est d'exploiter cet avantage.
5. **Transformation dans le data warehouse :** une architecture ELT permet aux analystes d'effectuer des transformations basées sur SQL dans un Warehouse élastique basé sur le Cloud. Les transformations basées sur SQL offrent également la possibilité de lancer des analyses à l'aide de modèles de données SQL prêts à l'emploi.

6. **Récupération en cas d'échec :** il est hors de question que vous perdiez définitivement vos données. Déterminez si les outils potentiels sont **idempotents** et effectuent une intégration additive nette.
7. **Sécurité et conformité :** deux domaines clés, tant en termes de protection des données que de perception par le public. Étudiez plus précisément comment les outils potentiels prennent en charge les domaines suivants :
    a. Conformité réglementaire
    b. Conservation limitée des données
    c. Contrôle des accès basé sur des rôles
    d. Blocage et hachage de colonnes

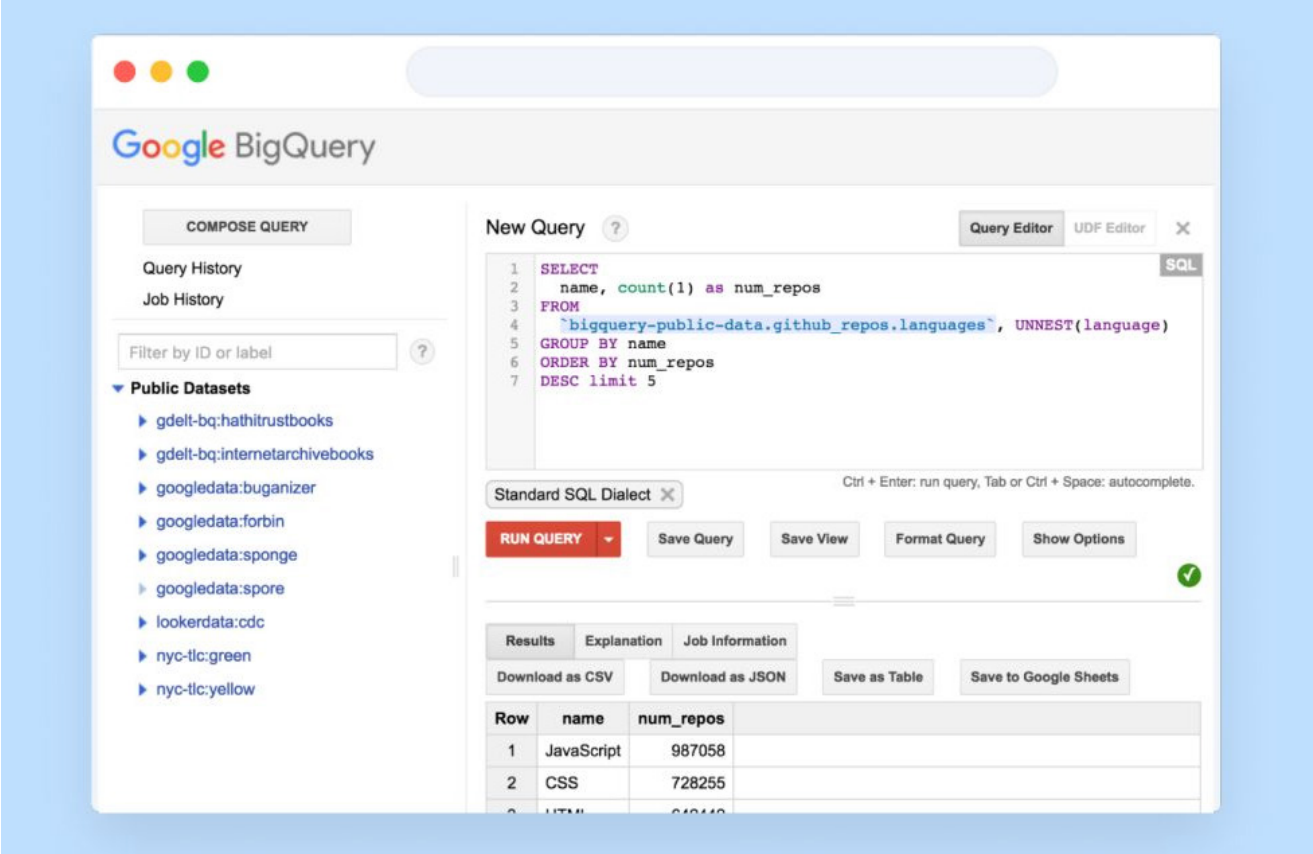


Fivetran offre une approche zero-touch de la Data Integration dans les data warehouses courants comme Snowflake.

Choisir le bon data warehouse

Votre data warehouse sera le dépôt d'enregistrement des données structurées de votre organisation. Les fonctionnalités et les compromis à consentir varient selon les data warehouses. Voici neuf critères à considérer :

1. **Stockage centralisé ou décentralisé des données :** le data warehouse stocke-t-il toutes ses données sur une seule machine, ou sont-elles réparties sur plusieurs machines, pour faire passer la redondance avant la performance ?
2. **Élasticité :** le data warehouse peut-il augmenter ou diminuer rapidement les ressources de calcul et de stockage ? Le calcul et le stockage sont-ils indépendants l'un de l'autre ou couplés ?
3. **Simultanéité :** le data warehouse peut-il augmenter ou diminuer rapidement les ressources de calcul et de stockage ? Le calcul et le stockage sont-ils indépendants l'un de l'autre ou couplés ?
4. **Performances des chargements et des requêtes :** à quelle vitesse pouvez-vous exécuter des chargements et des requêtes standard ?
5. **Gouvernance des données et gestion des métadonnées :** comment le data warehouse gère-t-il les autorisations et la conformité réglementaire ?
6. **Dialecte SQL :** quel dialecte SQL le data warehouse utilise-t-il ? Prend-il en charge les types de requêtes que vous souhaitez exécuter ? Vos analystes devront-ils adapter la syntaxe qu'ils utilisent actuellement ?
7. **Sauvegarde et récupération :** en cas de corruption ou de défaillance du data warehouse, pouvez-vous facilement revenir à un état antérieur ?
8. **Résilience et disponibilité :** qu'en est-il de la prévention des défaillances de base de données en premier lieu ?
9. **Sécurité :** le data warehouse respecte-t-il les bonnes pratiques actuelles de sécurité ?



Les data warehouses présentent des interfaces semblables à celles des bases de données opérationnelles.

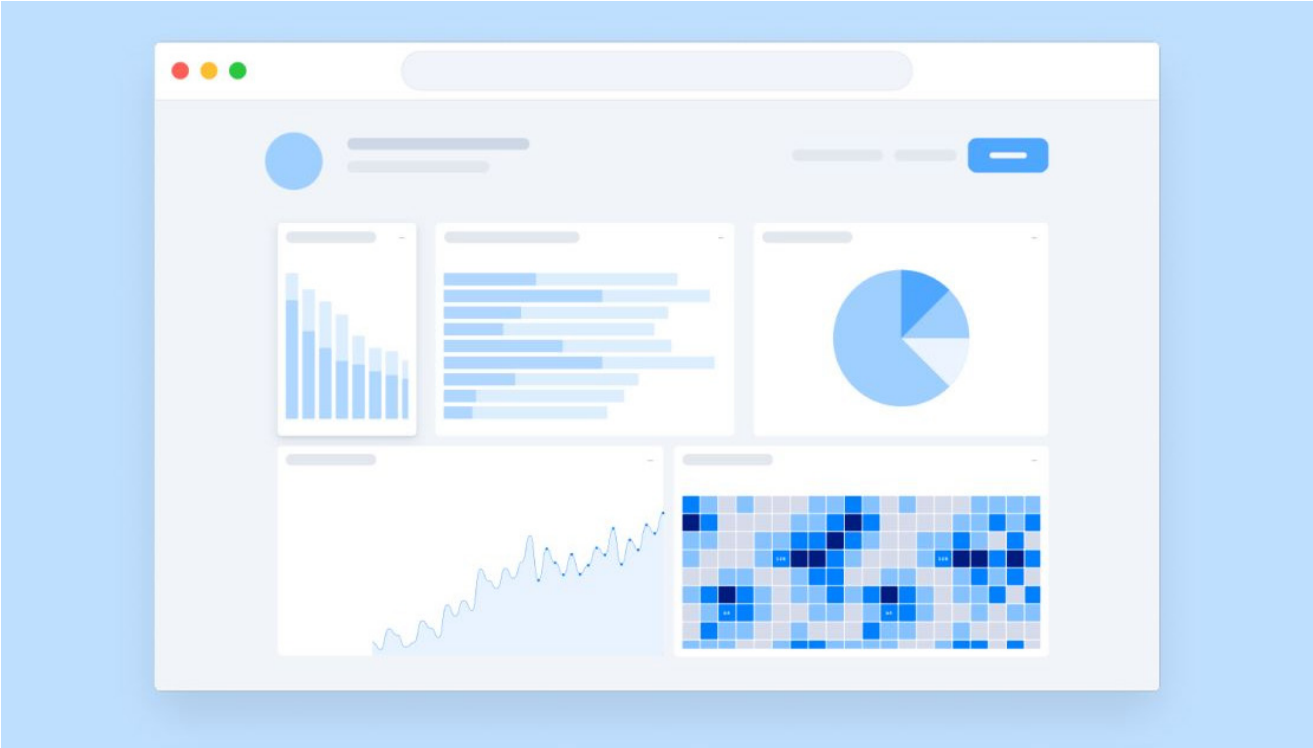
En tant que dépôt de vos données, le choix de votre data warehouse dans la composition de votre Data Stack sera financièrement déterminant. Les mises à jour et modifications de votre data warehouse impliquent des migrations de données très coûteuses.

Choisir le bon outil de Business Intelligence

Les outils de Business Intelligence vous permettent de créer facilement des rapports et des tableaux de bord. Cependant, chacun a ses points forts et ses faiblesses. Voici les facteurs clés dont vous devez tenir compte :

1. **Intégration transparente avec les data warehouses dans le Cloud :** est-il facile de connecter cet outil de BI au data warehouse de votre choix dans le Cloud ?
2. **Facilité d'utilisation et interfaces de type glisser-déposer :** la facilité d'utilisation est particulièrement importante pour mieux faire accepter les décisions fondées sur les données dans toute votre organisation.

3. **Reporting et notifications automatisés :** rédiger des rapports à la main peut devenir fastidieux pour les data scientists et les analystes. L'outil de BI vous permet-il de programmer la publication automatique des rapports ? Existe-t-il une fonctionnalité d'alerte des utilisateurs lorsque les données changent ?
4. **Possibilité d'effectuer des calculs et de générer des rapports ad hoc par ingestion et exportation de fichiers de données :** vos analystes et data scientists ont parfois besoin d'explorer les données sans avoir à passer par un data warehouse au préalable.
5. **Vitesse, performances et réactivité :** les questions de confort d'utilisation sont importantes, comme le chargement rapide des tableaux de bord et des visualisations.
6. **6. Couche de modélisation avec contrôle de version et mode de développement :** l'outil de BI offre-t-il à vos analystes un environnement collaboratif leur permettant de partager les modèles de données et le code ?
7. **7. Vaste bibliothèque de visualisations :** les diagrammes circulaires, les histogrammes, les courbes de tendance et autres visualisations de base ont leurs limites. L'outil de BI propose-t-il des visualisations plus spécialisées, comme des cartes thermiques ou des graphiques en radar ? Vous permet-il de créer vos propres visualisations personnalisées ?



Un outil de Business Intelligence peut prendre en charge les graphiques à barres, à colonnes et à secteurs, les courbes de tendance, les cartes thermiques et de nombreux autres types de visualisations.

Chapitre 6 : La Data Integration en six étapes

Concrètement, le parcours de Data Integration comporte six étapes :

1. Élimination des obstacles à une Modern Data Stack
2. Migration ou nouvelle instance
3. Évaluation des éléments de votre Modern Data Stack
4. Calcul du coût total de possession et du retour sur investissement
5. Définition des critères de réussite
6. Élaboration du Proof of Concept

Élimination des obstacles à une Modern Data Stack

La Modern Data Stack repose sur l'externalisation et l'automatisation de vos opérations de données. Cependant, il existe des raisons légitimes de ne pas faire appel à des fournisseurs tiers ou basés sur le Cloud.

La première et la plus évidente des raisons : la très petite taille de votre organisation ou le fait qu'elle fonctionne avec un faible volume ou niveau de complexité des données. Il se peut que vous n'ayez pas du tout d'opérations de données si vous êtes une toute petite start-up qui tente encore de trouver un produit adapté au marché. Il en va de même si vous n'utilisez qu'une ou deux applications, qu'il est peu probable que vous en adoptiez de nouvelles et que vos outils d'analyse intégrés pour chaque application vous suffisent.

Deuxième raison de ne pas faire l'acquisition d'une Modern Data Stack : le fait qu'elle ne réponde pas à certaines normes de performance ou de conformité réglementaire. Si vos opérations sont sensibles à une latence de quelques nanosecondes, mieux vaut éviter une infrastructure Cloud tierce et construire votre propre matériel.

La troisième raison concerne le cas où votre organisation produit ses propres logiciels spécialisés et utilise ou vend les données produites par ces logiciels. Et si vous proposez un service Web en streaming qui produit des téraoctets de données utilisateur chaque jour et soumet également des recommandations aux utilisateurs ? Même dans ce cas, votre organisation peut toujours externaliser les opérations de données pour les sources de données externes.

Dans les autres cas, si la taille et la maturité de votre organisation sont suffisantes pour tirer parti de l'Analytics et que des cycles d'actualisation des données de quelques minutes ou quelques heures sont acceptables, n'hésitez plus.

Migration ou nouvelle instance

Les fournisseurs de Data Integration doivent pouvoir migrer les données de l'ancienne infrastructure vers votre nouvelle Data Stack, mais la tâche est fastidieuse en raison du volume et de la complexité intrinsèque des données. La décision pour votre entreprise de migrer ou de créer une nouvelle instance à partir de zéro dépend fortement de l'importance qu'elle accorde aux données historiques.

La résiliation de contrats de produits ou services existants peut être coûteuse. Au-delà de l'aspect financier, la préférence pour certains outils et technologies ou leur maîtrise peut être une considération importante.

Assurez-vous que les solutions que vous envisagez sont compatibles avec les produits et services que vous souhaitez conserver.

Évaluation des éléments de votre Modern Data Stack

Vous aurez besoin d'un outil de Data Integration, d'un data warehouse, d'une plate-forme de Business Intelligence et d'une couche de transformation. Reportez-vous au chapitre précédent pour connaître les critères exacts sur lesquels vous appuyer pour évaluer vos choix et vous assurer que les technologies sont compatibles entre elles.

Calcul du coût total de possession et du retour sur investissement

La Modern Data Stack promet des économies substantielles de temps, de talents et d'argent. Comparez votre flux de travail de Data Integration existant avec les alternatives.

Calculez le coût de votre Data Pipeline actuel. Le principal facteur est probablement le temps que votre équipe de données a consacré à la construction et à la maintenance du Data Pipeline. Un audit méticuleux des outils que vous utilisez pour la gestion de projet peut être nécessaire.

Vous devrez également tenir compte du prix courant des outils et des technologies utilisés. Enfin, vous devrez considérer les coûts d'opportunité liés aux défaillances, arrêts et temps d'immobilisation. Intégrez aussi les coûts de votre data warehouse et de votre outil de BI.

En second lieu, vous devrez évaluer les avantages de la solution de remplacement envisagée. Si certains sont sans doute difficiles à mesurer (comme l'amélioration potentielle de la motivation des analystes), d'autres, tels que les gains de temps et d'argent, sont aisément quantifiables.

Définition des critères de réussite

Une solution de Data Integration automatisée peut remplir un certain nombre d'objectifs. Définissez vos critères de réussite à partir de ces éléments :

1. **Économies de temps, d'argent et de main-d'œuvre :** une Modern Data Stack doit permettre de réduire considérablement vos coûts d'ingénierie des données en éliminant la nécessité de construire et de maintenir des connecteurs de données. Les économies de main-d'œuvre peuvent représenter des centaines d'heures d'ingénierie par semaine, soit des sommes considérables. Vous pouvez **utiliser notre calculateur** pour obtenir une première estimation.
2. **Capacités étendues :** une Modern Data Stack (MDS) doit élargir les capacités de votre équipe chargée des données en mettant à disposition davantage de sources de données sans travail supplémentaire.
3. **Exécution réussie de nouveaux projets de données, tels que les modèles d'attribution des clients :** avec davantage de temps et de sources de données, votre équipe peut construire de nouveaux modèles de données, y compris ceux qui suivent les mêmes entités sur plusieurs sources de données.
4. **Réduction du délai d'exécution des rapports :** une Modern Data Stack doit réduire considérablement le délai d'exécution des rapports et permettre ainsi aux décideurs clés de se maintenir à jour.
5. **Réduction des temps d'arrêt de l'infrastructure de données :** une Modern Data Stack doit considérablement améliorer la fiabilité et quasiment éliminer votre travail de maintenance.

6. **Utilisation accrue de la Business Intelligence :** en combinant l'intégration automatisée des données avec un outil de BI moderne et intuitif, une Modern Data Stack doit favoriser l'accès aux données, leur connaissance et leur utilisation dans toute l'entreprise.
7. **Nouvelles métriques disponibles et exploitables :** dotée de sources de données supplémentaires et d'un outil de BI facile à utiliser, une Modern Data Stack doit permettre d'obtenir de nouvelles métriques et de nouveaux indicateurs clés de performance pour la prise de décision.

Élaboration du Proof of Concept

Après avoir limité votre recherche à quelques outils possibles et déterminé les critères de réussite, testez les produits sur des projets à faible enjeu. La plupart des produits proposent des essais gratuits de quelques semaines.

Configurez les connecteurs entre vos sources de données et vos data warehouses, puis mesurez le temps et les efforts requis pour synchroniser les données. Effectuez ensuite quelques transformations de base. Aménagez une plage d'essai pour votre équipe et encouragez-la à tester les performances du système de toutes les façons possibles. Comparez les résultats de votre essai à vos standards de réussite.

Même si vous avez éliminé les difficultés techniques, d'autres obstacles peuvent entraver l'adoption d'une Modern Data Stack. Votre équipe chargée des données peut manquer de financement, d'effectifs ou d'expertise. Les data engineers peuvent vouloir protéger les systèmes qu'ils ont construits. La direction peut avoir des difficultés à percevoir l'intérêt d'un redimensionnement rapide de la Data Integration. Il est important d'obtenir l'adhésion d'une personne habilitée à acheter les outils et les technologies nécessaires, ainsi que de cultiver un état d'esprit moderne à l'égard des données dès le début de votre parcours.

C'est exactement ce que peut faire une démonstration de produit minimum viable (MVP) soigneusement élaborée qui prouve la valeur de la Modern Data Stack sur une seule source de données, un seul rapport ou un seul scénario test.

Chapitre 7 : Comment continuer à moderniser votre Analytics

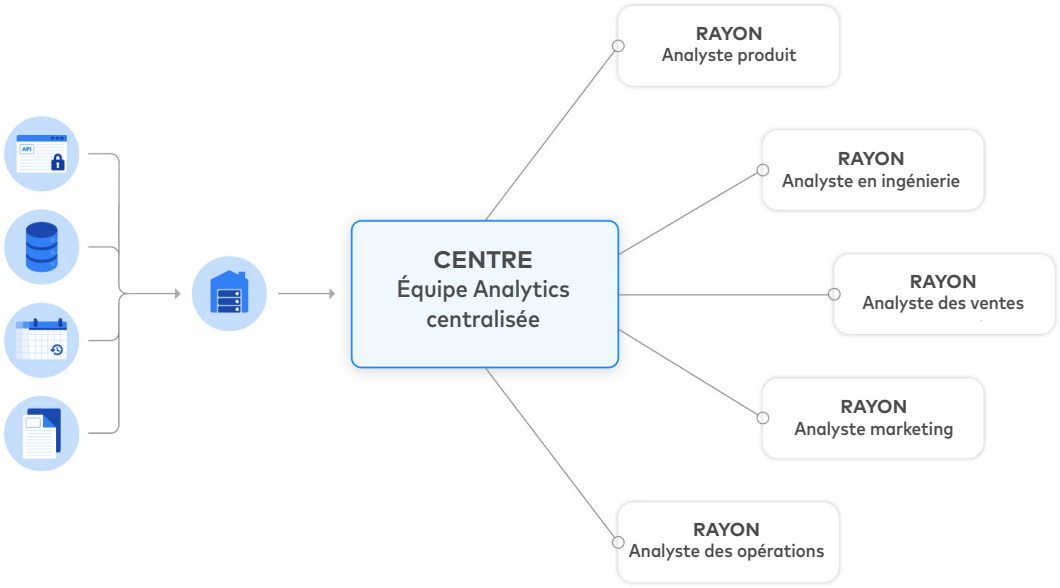
Une Modern Data Stack n'est que la première pierre dans la construction d'une opération de données mature. Munie des bons outils, votre équipe dédiée sera en mesure d'extraire, de charger, de modéliser et de transformer vos données. Toutefois, des changements organisationnels progressifs seront nécessaires pour développer vos nouvelles capacités et continuer à en faire bon usage. Ces changements sont les suivants :

1. Faire évoluer votre organisation analytique
2. Établir des normes de gouvernance des données
3. Penser produit
4. Promouvoir la culture des données
5. Construire une architecture de données robuste
6. Embaucher des data scientists

Faire évoluer votre organisation analytique

Une fois votre Modern Data Stack en place, la première chose à faire est d'étoffer votre équipe chargée des données. Le fer de lance de vos équipes Analytics n'est autre que les analystes, dont l'expertise permet de créer des modèles de données, des tableaux de bord et des rapports pour aider votre entreprise à prendre des décisions.

À mesure que vos besoins en données augmentent et deviennent plus complexes, vous devrez gonfler les rangs de vos analystes. Il existe autant d'arguments en faveur des équipes de données centralisées que décentralisées, le bon compromis consistant à associer les deux dans le cadre d'un **modèle en étoile.**



Votre « centre » est une équipe centralisée responsable du processus global de Data Integration, qui crée et maintient les modèles de données, les rapports et les visualisations moins spécialisés utilisés par la direction de l'entreprise et les contributeurs individuels. Les « rayons » sont de petites équipes d'analystes fonctionnellement alignées, qui sont intégrées à des départements spécifiques et possèdent une expertise dans des domaines tels que les ventes, la finance, etc. Le centre et les rayons doivent rendre compte à la direction de l'entreprise. On veillera à affecter un des postes de la direction à cette fonction (Chief Analytics Officer, Chief Data Officer ou équivalent).

Établir des normes de gouvernance des données

La **gouvernance des données** est l'autre aspect primordial du processus. La propriété des éléments de données implique également de documenter et de cataloguer soigneusement tous ces éléments. Ces opérations gagneront en importance à mesure que vous ingérez des données et que vous développez davantage de produits ou de branches d'activité. Sans gestion appropriée, un « marécage de données » peut rendre les données inexploitables. Le respect des normes réglementaires (et des considérations éthiques élémentaires) devient dès lors difficile, voire impossible.

Pour éviter ce problème, envisagez de recourir à un outil de catalogue de données Cloud. Cela vous aidera à effectuer les opérations suivantes :

- Documenter tous les modèles, tables et champs. La tâche peut s'avérer fastidieuse si vous disposez de nombreuses sources de données. Une alternative consiste à élaborer soigneusement un schéma dimensionnel, à savoir un modèle de données simplifié englobant toutes les principales opérations.
- Déterminer les métriques dont vous avez besoin et leur provenance.
- Noter la fréquence à laquelle vous devez actualiser les données.
- Prévoir la résolution des problèmes d'intégrité des données.
- Identifier les véritables propriétaires des données pour les différents modèles au sein de l'organisation.
- Attribuer la propriété et créer des incitations pour maintenir le bon fonctionnement du système.

La mise en place de votre Modern Data Stack est le meilleur moment pour réaliser ce travail, car vous devrez de toute façon faire l'inventaire de tous les éléments de données.

Intégrez le contrôle de la gouvernance des données dès le départ afin d'instaurer la confiance. Sans une provenance claire de chaque modèle de données, il sera difficile pour les utilisateurs finaux de comprendre comment les métriques sont déterminées et de trancher les récits contradictoires.

Penser produit

À mesure que votre équipe chargée des données s'implique davantage dans la visualisation des données et l'aide à la décision, vous devrez faire un effort concerté pour intégrer la pensée produit au travail d'Analytics. Vous aurez besoin d'un nouveau rôle, celui de gestionnaire de produits de données, pour diriger la création des éléments de données.

En bref, penser produit consiste à comprendre vos utilisateurs (c'est-à-dire les contributeurs individuels et les dirigeants de votre entreprise) et à produire et affiner les produits de manière rapide et itérative en réponse à l'évolution des conditions. Plus précisément, le processus produit pour les éléments de données se déroule de la façon suivante :

Identification

- Compréhension des utilisateurs
- Recueil des besoins

Conception

- Définition de la portée
- Gestion des attentes

Développement

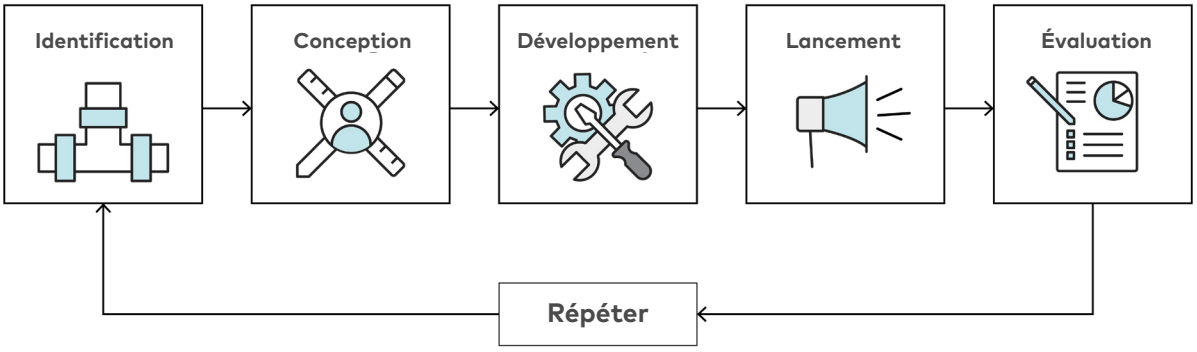
- Prototypage rapide
- Mise en production

Lancement

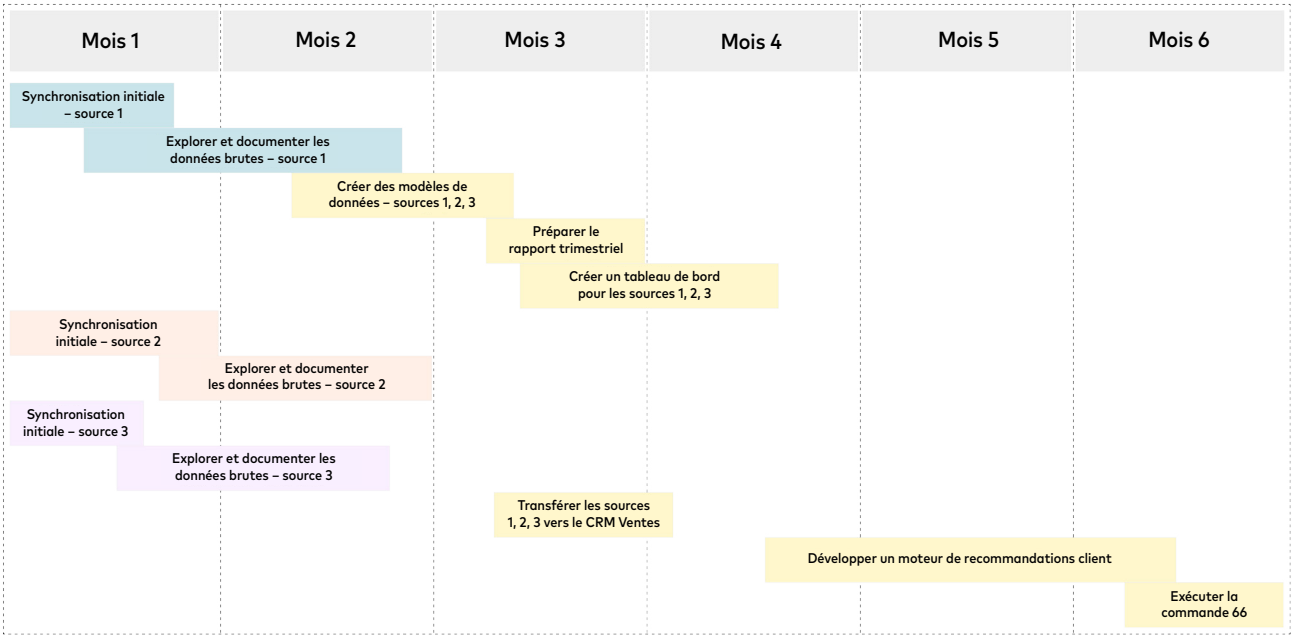
- Marketing et déploiement du produit
- Formation des utilisateurs aux heures de bureau et par communication interne
- Soutien de l'adoption, notamment par le biais du libre-service chaque fois que possible

Évaluation

- Évaluation par rapport aux attentes et aux KPI



Vous devrez établir une feuille de route sur la manière dont les éléments de données amélioreront la prise de décision dans votre entreprise. Dressez la liste des étapes et des objectifs finaux, ainsi que des informations et des connaissances nécessaires pour les réaliser. Vous pouvez présenter votre feuille de route à votre direction, en proposant de l'intégrer à la stratégie de l'entreprise.



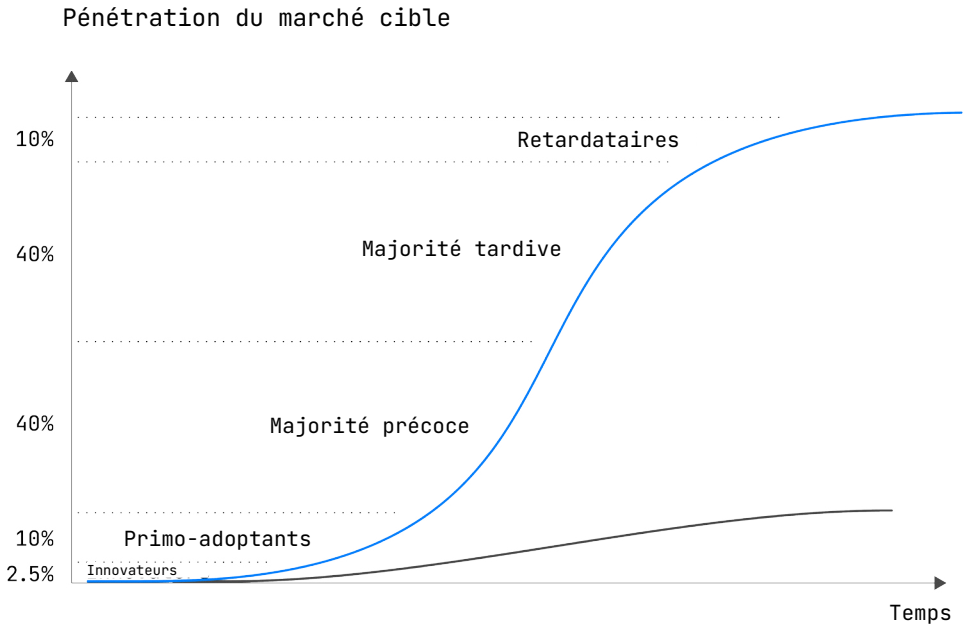
Vous devez tenir compte de certains points particuliers lorsque vous communiquez un programme axé sur les données. Les moins expérimentés dans l'utilisation des outils et métriques appropriés d'analyse des données peuvent se montrer sceptiques. Des préjugés dystopiques sur les possibilités d'utilisation douteuse ou contraire à l'éthique des données peuvent être exprimés.

Il est important d'anticiper ces risques et de communiquer sur les avantages de l'orientation données en termes clairs qui répondent aux priorités de votre public, pas seulement aux vôtres. Évitez le jargon à la mode et les détails trop techniques ; mettez plutôt en lumière la manière dont les résultats souhaités, fondés sur les données, seront atteints. Il est essentiel de partager un récit autour des données et de le ponctuer d'exemples concrets et d'éléments probants. Vous devrez combiner de manière transparente la stratégie, les visualisations et le narratif. Vous inciterez ainsi le personnel à se familiariser avec les données et à les utiliser.

Promouvoir la culture des données

Alors que votre équipe chargée des données soutient la pensée produit, vous devrez également promouvoir la culture des données auprès de la direction et des contributeurs individuels. L'expression « culture citoyenne » est parfois employée pour décrire la décentralisation de la prise de décision. Il n'est pas raisonnable d'exiger des analystes et autres professionnels des données qu'ils se fassent les interprètes des décideurs.

Il peut être utile de considérer la culture des données à la lumière de la **théorie de la diffusion des innovations.** Imaginez une courbe en S dans laquelle 2,5 % des personnes sont des innovateurs et des pionniers, 13,5 % des primo-adoptants et 34 % la majorité précoce. La principale difficulté au départ est de trouver une personne enthousiaste et techniquement compétente en position d'influence, capable de convaincre son équipe en votre nom. À mesure que les équipes se familiarisent avec les données et deviennent plus compétentes, vos efforts sont appelés à faire boule de neige et à monter en puissance.

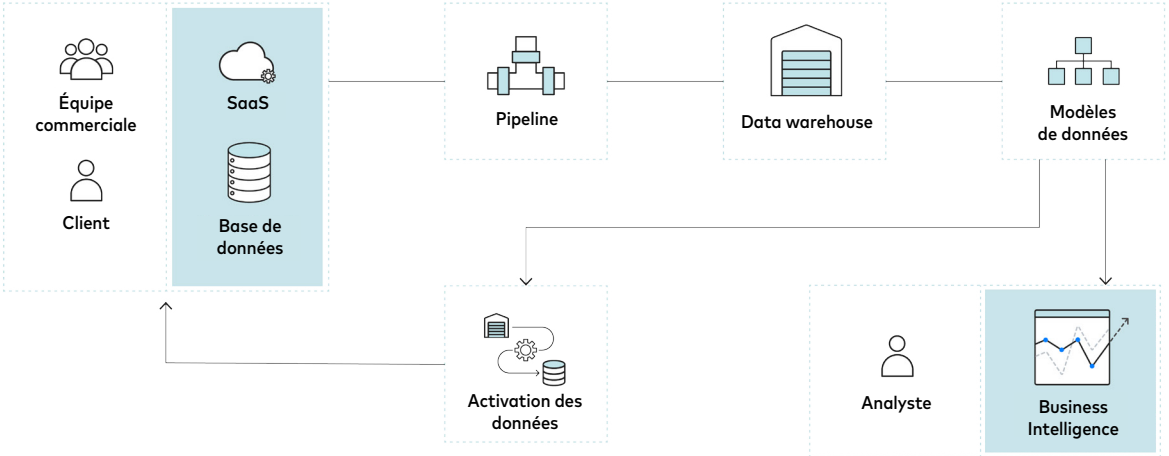


À l'avenir, en faisant de la culture des données un critère essentiel d'embauche, il sera plus facile d'améliorer les capacités de vos équipes en matière de données. N'oubliez pas que la base de référence n'est pas l'écriture de requêtes SQL ou la création de nouveaux modèles, mais une capacité générale à interpréter des graphiques et des tableaux et à prendre des décisions en conséquence.

Construire une architecture de données robuste

L'architecture des données désigne l'ensemble des outils et des processus employés pour la Data Integration. À mesure que vos opérations de données gagnent en maturité, vous ajouterez de nouveaux outils et technologies, utiliserez d'autres fonctionnalités des outils existants, apporterez des changements organisationnels et créerez de nouveaux flux de travail. Il sera d'autant plus important de « boucler la boucle » en activant les données analytiques et en transformant les insights en décisions commerciales et en actions percutantes. Plus précisément, vous devrez tenir compte de ces éléments :

- **Redimensionner la capacité pour faciliter l'acceptation de nouvelles sources de données :** assurez-vous que votre équipe chargée des données et les outils qu'elle utilise sont capables d'étendre vos efforts de Data Integration et d'Analytics à une grande variété de sources de données. Il s'agit de trouver un moyen évolutif de produire des modèles de données et de s'assurer que votre équipe chargée des données possède une expertise dans un certain nombre de fonctions commerciales différentes.
- **Rapports automatisés :** à mesure que votre organisation se développe, vous ne devrez plus compter sur les analystes pour produire des rapports ou des tableaux de bord sur commande, alors que de nombreuses plates-formes de BI modernes peuvent le faire de manière planifiée.
- **Contrôle automatisé de la Data Integration :** à mesure que le nombre de vos utilisateurs et de vos sources de données augmente, vous devrez disposer d'un moyen d'automatiser le contrôle des programmations, l'attribution des autorisations et la gestion de vos Data Pipelines.
- **Activation des données pour une utilisation en production :** vous aurez besoin d'une combinaison d'outils prêts à l'emploi et d'expertise en ingénierie des données pour acheminer les données analytiques dans les systèmes opérationnels et de production.



Embaucher des data scientists

Félicitations ! Vous avez atteint le sommet de la hiérarchie des besoins en matière de données que nous avons abordée au chapitre 1. Vous êtes maintenant prêt à recruter des data scientists, et même à commencer à exploiter l'IA et le machine learning pour obtenir des quantités exponentielles d'analyses de données et des insights de valeur.

Les data scientists combinent une expertise en statistiques appliquées et en algèbre linéaire avec suffisamment de compétences en ingénierie pour prototyper des modèles de machine learning et les mettre en production avec l'aide de data engineers.

De nombreuses organisations brûlent les étapes et engagent des data scientists pour réaliser le travail des analystes ou des data engineers avant de mettre en place une véritable hiérarchie des données. C'est une erreur. Il est préférable d'embaucher des data scientists lorsque votre infrastructure de données arrive à maturité et que votre entreprise a répondu aux besoins plus essentiels en matière de données.

Confiez à des data scientists le soin de concevoir, de construire, de tester et d'ajuster des modèles de machine learning à mesure que votre entreprise s'oriente vers la modélisation prédictive et l'intelligence artificielle.

Vous devriez à présent avoir une idée précise de la Modern Data Stack et de l'état d'esprit nécessaire pour l'intégrer dans votre organisation, entouré d'équipes motivées et d'un leadership solidaire. Nous vous souhaitons bonne chance dans votre parcours, et vous remercions de nous avoir lus!

Fivetran peut vous aider à franchir la première étape de votre parcours de modernisation de l'Analytics.

Demandez une démonstration sur https://get.fivetran.com/demo ou commencez un essai gratuit sur https://fivetran.com/signup dès aujourd'hui.

Fivetran automatise le mouvement des données à la fois depuis, vers et à travers les plateformes de données cloud. Fivetran automatise les parties les plus laborieuses du processus d'ELT, de l'extraction à la transformation, afin que les ingénieurs data puissent se concentrer sur des projets à plus fort potentiel, tout en ayant l'esprit tranquille. Grâce à un taux de disponibilité de 99,9 % et à des pipelines à rétablissement automatique, Fivetran permet à des centaines de grandes marques du monde entier, dont Autodesk, Conagra Brands, JetBlue, Lionsgate, Morgan Stanley et Ziff Davis, d'accélérer leurs décisions fondées sur la data et de soutenir leur croissance. Fivetran est basée à Oakland, en Californie, et possède des bureaux dans le monde entier. Pour plus d'informations, visitez **Fivetran.com.**