Guide essentiel de l'intégration de données

Comment prospérer à l'ère des données

Par Charles Wang

Sujets traités dans ce guide :

- Comment l'intégration de données alimente l'analyse
- ETL, ELT et intégration de données automatisée
- Avantages de l'intégration de données automatisée
- Comment évaluer les fournisseurs de services d'intégration de données

Fivetran

fivetran.com →

Sommaire

À propos de cet ouvrage

Objectifs

Cet ouvrage démontre l'utilité de l'intégration de données pour votre entreprise, décrit et évalue les différentes approches en la matière et explique comment mettre en œuvre cette technologie. S'il n'est pas nécessaire de lire ce document du début à la fin, il est structuré pour être aisément assimilé de cette manière.

L'**intégration de données** est le processus consistant à gérer et à centraliser les flux de données depuis plusieurs sources, afin de les utiliser pour orienter la prise de décision. Dans ce contexte, l'interprétation pratique des données est appelée **analyse de données**. Comme nous le verrons, la qualité de votre programme d'analyse de données est étroitement liée à celle de votre technologie d'intégration de données.

L'intégration de données permet à votre entreprise de conserver l'ensemble de ses données au sein d'un environnement unique, afin d'offrir à votre équipe une vue complète des opérations métier et interactions client. La centralisation et la mise à disposition des données favorisent leur culture à l'échelle de l'entreprise, vous permettant de déceler des opportunités dissimulées, d'améliorer les performances et de stimuler l'innovation.

Dans ce guide, nous traiterons les points suivants :

- Qu'est-ce que l'intégration de données et pourquoi est-elle essentielle
- L'approche traditionnelle de l'intégration de données, appelée ETL (« Extract-Transform-Load », Extraire-Transformer-Charger)
- L'approche moderne axée sur le cloud, appelée ELT (« Extract-Load-Transform », Extraire-Charger-Transformer)
- Avantages liés à l'automatisation de l'intégration de données
- Comment évaluer et adopter des outils d'intégration de données

Public ciblé

Pour tirer pleinement parti de ce guide, vous devez être familiarisé avec les processus d'ingénierie des données, d'entreposage de données, d'analyse de données, d'informatique décisionnelle (BI), de visualisation de données et les concepts associés. Nous partons du principe que votre entreprise utilise des systèmes, applications et autres

outils opérationnels produisant des données numériques et que certaines de ses activités sont déjà basées sur le cloud. Nous supposons également que votre fonction — analyste, ingénieur des données, expert des données ou responsable d'un de ces postes — vous permet d'influencer le choix ou de décider des outils utilisés par votre entreprise.

Icônes utilisées dans ce guide

Au fil de votre lecture, vous rencontrerez des icônes désignant les conseils, avertissements, points à retenir et études de cas.

CONSEIL : *Conseils pratiques concernant l'intégration et l'analyse de données*

ATTENTION : *Exemples de mauvaises pratiques techniques ou liées aux données*

N'OUBLIEZ PAS : *Points importants à retenir*

ÉTUDE DE CAS : *Cas de réussite réels en matière d'intégration de données*

Au-delà de ce guide

Si ce guide vous a été utile et que vous souhaitez en apprendre davantage, visitez la page *fivetran.com/blog* sur laquelle nous publions les nouveaux contenus en matière d'ingénierie et d'analyse de données. Notre documentation, disponible sur la page *fivetran.com/docs*, est également une ressource pertinente ; elle détaille le fonctionnement de l'intégration de données pour des sources et destinations spécifiques.

Chapitre 1 : Intégration et analyse de données

DANS CE CHAPITRE :

- Présentation historique de l'analyse de données
- Valeur de l'analyse de données pour vous et votre entreprise
- Un obstacle majeur à l'analyse de données : l'intégration de données

Historique de l'analyse de données

L'analyse de données est bien antérieure aux processus de collecte de données modernes. Florence Nightingale utilisait des digrammes circulaires pour identifier et réduire les causes de mortalité dans les hôpitaux lors de la guerre de Crimée (Figure 1.0). William Sealy Gosset, brasseur en chef chez Guinness, a développé le test t (de Student) pour garantir la qualité de la bière. Depuis des temps très reculés, l'homme exploite les chiffres pour parvenir à des conclusions pratiques et éclairées.

Depuis lors, les statistiques ont continué d'évoluer en tant que science, à l'instar des outils et méthodes d'analyse de données. L'essor de l'informatique moderne, et en particulier d'Internet, a permis la collecte et l'analyse de données à une échelle bien plus vaste que les approches possibles à l'aide de crayons, de papier et de tabulatrices.

Objectifs de l'analyse de données

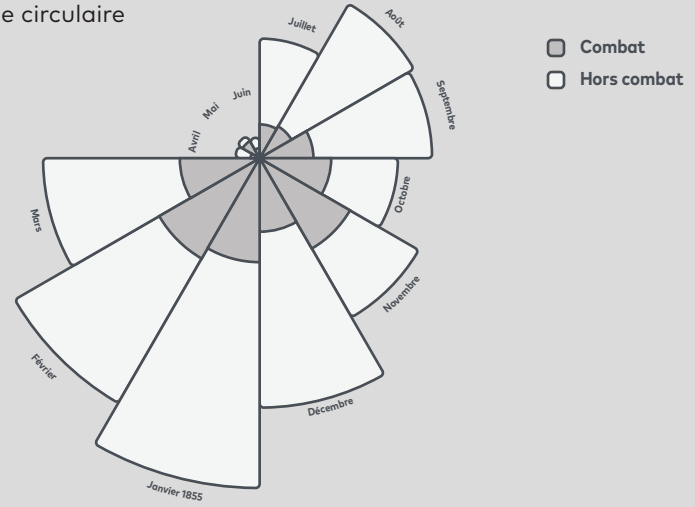
L'analyse de données offre une valeur concurrentielle dans plusieurs domaines. Elle peut être utilisée pour améliorer l'acquisition, la rétention et la fidélité des clients, ainsi que pour identifier de nouvelles opportunités produit et améliorer les opportunités

existantes. En optimisant le processus de décision organisationnel, l'analyse de données peut restituer plusieurs fois son coût en retour sur investissement.

D'une manière générale, vous pouvez utiliser l'analyse de données aux fins suivantes :

1. **Génération de rapports ad hoc.** Les parties prenantes et décisionnaires clés peuvent parfois nécessiter des réponses très spécifiques sur une base ponctuelle ou occasionnelle.
2. **Informatique décisionnelle.** Souvent utilisé comme équivalent de l'« analyse de données », l'informatique décisionnelle (BI pour « Business Intelligence ») se rapporte à l'utilisation de visualisations et modèles de données visant à identifier les opportunités, tout en orientant les décisions et stratégies métier. Cette approche prend généralement la forme de rapports réguliers et cohérents et de tableaux de bord actualisés.
3. **Données en tant que produit.** Les données collectées ou produites par votre entreprise peuvent être mises à la disposition de tiers sous forme de tableaux de bord intégrés, de flux de données, de recommandations ou d'autres produits.
4. **Intelligence artificielle/apprentissage automatique.** Le summum en matière d'analyse de données consiste à créer des produits et systèmes utilisant une modélisation prédictive pour automatiser des décisions et processus critiques.

Figure 1.0
Diagramme circulaire



Source : « Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army, » par Florence Nightingale. Londres : Harrison & Sons, 1858.

Au niveau organisationnel, l'analyse de données facilite la poursuite des objectifs suivants :

1. **Démocratisation de l'accès aux données/de la culture des données.** Plus le

nombre d'employés utilisant des données pour prendre des décisions est important, plus votre entreprise réagit intelligemment à l'évolution de la situation. Avec des outils de BI adaptés, même les collaborateurs sans compétences techniques peuvent prendre des décisions basées sur les données. Bien entendu, cela implique que vous accordiez à votre équipe un niveau de confiance et de latitude significatif.

2. **Amélioration de vos produits et services.** Les perspectives issues de l'analyse de données vous aideront à améliorer vos offres tout en augmentant la transparence et la gamme de rapports pour vos clients.
3. **Maintenir la compétitivité de votre entreprise.** La culture des données vous permet de tirer pleinement parti de ressources limitées et de dévoiler des opportunités autrement invisibles.

La connaissance est synonyme de pouvoir, et il est toujours avantageux d'en savoir plus que la concurrence.

CONSEIL : *Il peut être pertinent d'envisager l'ensemble des activités liées aux données sous forme de niveaux dans une hiérarchie de besoins au sein de laquelle la satisfaction des besoins fondamentaux permet la poursuite des besoins de niveau supérieur (Figure 1.1).*

Le besoin le plus élémentaire est la collecte et le stockage de données brutes, c'est-à-dire l'intégration de données. Une fois ce besoin satisfait, il est plus facile de répondre aux besoins intermédiaires, tels que l'analyse et la modélisation prédictive des données. Cela permet à votre entreprise de développer une culture orientée données, au sein de laquelle chaque employé a accès aux données dont il a besoin pour prendre des décisions mieux avisées.

Au sommet de la hiérarchie, les données sont utilisées pour développer des modèles d'apprentissage automatique et l'intelligence artificielle, permettant l'automatisation des flux de travail et des processus décisionnels au sein de votre entreprise, ainsi que la création de produits « intelligents » destinés aux clients.

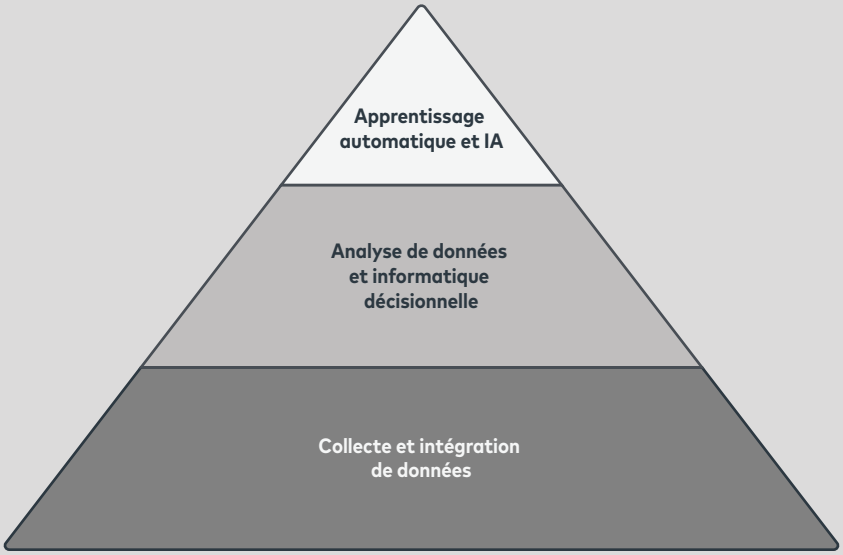
Un obstacle majeur à l'analyse de données : l'intégration de données

Un système de registre central vous offre les avantages suivants :

1. Vous bénéficiez d'une perspective globale des opérations de l'entreprise et d'une visibilité sur les interactions entre chaque élément, au lieu de représentations compartimentées et isolées.

2. Vous pouvez mettre les fichiers en correspondance et suivre les mêmes entités (client, partenaire, etc.) à travers les différentes phases de leur cycle de vie.
3. Vous pouvez exécuter les processus d'analyse dans un environnement séparé des systèmes opérationnels, empêchant ainsi les requêtes d'interférer avec vos opérations.
4. Vous disposez d'un contrôle granulaire sur les accès et autorisations, garantissant ainsi que votre équipe reçoit les informations dont elle a besoin pour remplir sa fonction sans compromettre les systèmes sensibles.

Figure 1.1
Hiérarchie des besoins liés aux données



La création de ce registre de données central peut être une tâche herculéenne. Chaque source nécessite des procédures et outils distincts pour ingérer, nettoyer et modéliser ses données. Ce défi a été amplifié par la prolifération récente des applications et services basés sur le cloud. L'apparition d'appareils et de capteurs connectés (c.-à-d. l'Internet des objets) a également contribué à une explosion des volumes de données (Figure 1.2). Depuis 2013, 90 % des données mondiales sont créées au cours des deux dernières années.[1]

D'où proviennent les données ?

Les données peuvent être issues des sources suivantes :

1. **Entrées de capteur**, telles que les lectures de codes aux caisses

1 sciencedaily.com/releases/2013/05/130522085217.htm

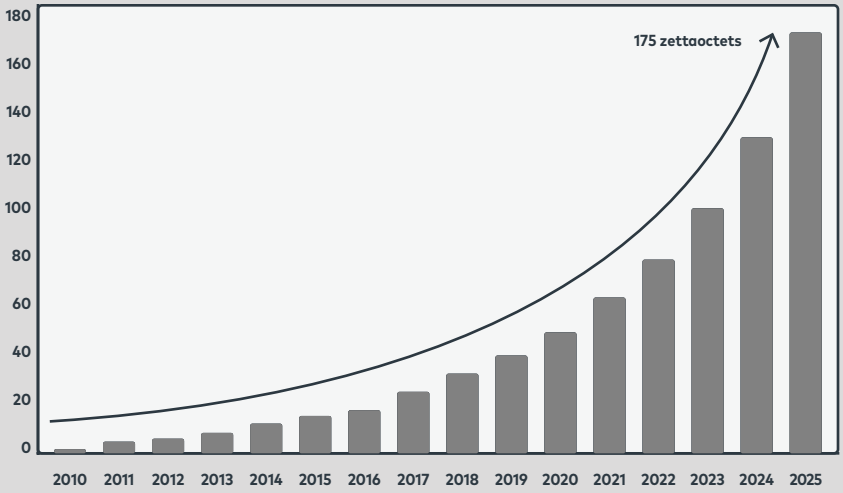
2. **Saisies de données manuelles,** telles que les formulaires collectés par le Bureau du recensement
3. **Documents et contenus numériques**, tels que les publications sur les réseaux sociaux
4. **Activités numériques** consignées par des déclencheurs logiciels, tels que les clics sur un site ou une application Web

Les données provenant des sources ci-dessus sont généralement stockées dans des fichiers numériques cloud et des bases de données opérationnelles, puis délivrées à un utilisateur final sous forme de :

1. Flux d'API
2. Fichiers
3. Journaux de base de données et résultats de requête
4. Suivi d'événements

Les flux d'API permettent aux applications de communiquer entre elles, souvent en échangeant des données au format JSON, ou encore XML. La plupart des entreprises utilisent un large éventail d'applications pour traiter les opérations telles que la gestion des relations client, la facturation et le service client, entre autres. Les API permettent l'ingestion des données et l'interopérabilité des applications logicielles.

Figure 1.2
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

Les **fichiers** de données, par exemple au format CSV, XLSX ou TSV, peuvent être issus de diverses activités au sein de l'entreprise, de la collecte de données manuelle aux calculs ad hoc.

Les **journaux de base de données et les résultats de requête** sont générés par des bases de données opérationnelles mises à jour en temps réel. Elles prennent en charge les interactions quotidiennes de tous les composants, des capteurs aux logiciels. Un site d'e-commerce peut par exemple utiliser une base de données opérationnelle pour consigner les achats, les référencements et les profils utilisateur.

Le **suivi des événements** intervient par le biais de fragments de code déclenchés par l'utilisateur intégrés aux pages et applications Web. Un outil de base peut enregistrer les clics dans une application, une version plus avancée suivre la position du curseur et une solution de pointe utiliser la caméra d'un ordinateur portable pour suivre le regard de l'utilisateur. Les données de suivi des événements génèrent un registre granulaire des interactions utilisateur avec un site ou une application Web et sont particulièrement utiles à des fins de recherche en matière d'interface/expérience utilisateur. L'une des formes les plus courantes est le **webhook**, qui est intégré aux applications Web et envoyé en HTML au lieu d'être formaté en XML ou JSON.

Comme vous pouvez l'imaginer, la multiplicité des sources et formats de données crée des défis considérables pour les ingénieurs tentant d'intégrer et de normaliser les flux de données.

CONSEIL : *Il se peut que vous rencontriez le terme « intégrité des données » dans divers domaines de traitement des données, notamment l'intégration de données. L'intégrité des données se rapporte à l'exhaustivité, à la précision et à la cohérence des données durant toutes les phases de leur utilisation. Les violations d'intégrité des données comprennent les erreurs de saisie ou de formatage des données, les duplications, omissions et associations incorrectes entre les tables.*

Données SaaS : des défis et opportunités croissants

À l'ère du cloud, les applications SaaS sont devenues l'une des principales sources de données d'entreprise et couvrent une multitude d'opérations et de secteurs : marketing, traitement des paiements, gestion des relations client, e-commerce, gestion des projets d'ingénierie, et bien d'autres. Elles fournissent des services et opérations sophistiqués, éliminant le besoin de développement d'outils en interne ou d'investissement massif en main-d'œuvre pour réaliser les mêmes tâches manuellement.

Les applications SaaS consignent généralement les actions réalisées par les utilisateurs, offrant ainsi aux entreprises une vision granulaire de leurs opérations,

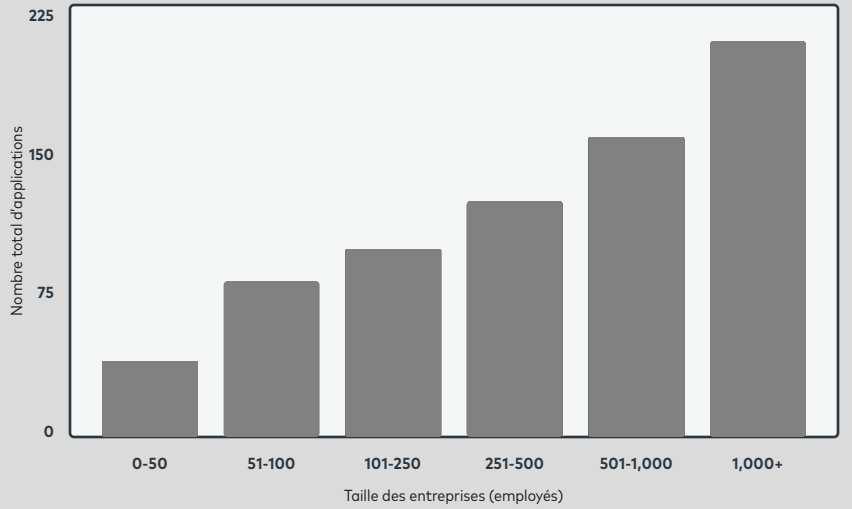
ce qui leur permet de déduire des schémas et relations de causalité. D'une manière générale, plus le nombre de facettes de votre activité que vous pouvez quantifier et analyser est important, plus votre entreprise est compétitive.

Toutefois, les volumes massifs de données posent un défi majeur en matière d'intégration. Les entreprises actuelles utilisent en moyenne plus de 100 applications (Figure 1.3), et à cette échelle, l'intégration de données manuelle est quasiment impossible. Comme nous le verrons, nombreuses sont les entreprises qui développent encore des logiciels et infrastructures personnalisés pour intégrer les données, mais cette approche devient ingérable lorsque des dizaines de sources génèrent d'importants volumes en continu.

Même à plus petite échelle, la charge de travail imposée par la mise en œuvre et la gestion d'un pipeline de données complexe peut entraver les efforts d'analyse, car ces opérations chronophages détournent les analystes, experts et ingénieurs des données des autres activités.

Heureusement, la technologie cloud offre une solution à cette problématique. Les outils de pipeline de données modernes, entrepôts de données et plateformes de BI sont des applications cloud en tant que telles qui ont proliféré aux côtés de cette technologie. Elles éliminent efficacement le besoin de développement manuel d'outils et de solutions personnalisés en interne pour l'intégration et l'analyse de données.

Figure 1.3
Nombre d'applications par entreprise



Source : rapport annuel sur les tendances SaaS 2019 de Blissfully

N'OUBLIEZ PAS : *Les fonctionnalités d'analyse permettent aux entreprises d'opérer au meilleur de leurs capacités, mais pour que vos analyses soient performantes et exhaustives, vos données doivent être accessibles dans un environnement centralisé. Un registre de données central vous offre les avantages suivants :*

- *Développement d'une vision globale et exhaustive des opérations à l'échelle de l'entreprise*
- *Mise en correspondance des fichiers et suivi des mêmes entités à travers différentes sources de données*
- *Séparation entre processus d'analyse et systèmes opérationnels*
- *Contrôle des accès et autorisations*

L'intégration des données, consistant à centraliser les données et à les rendre accessibles, est un processus à la fois essentiel et extrêmement difficile à mettre en œuvre. Ne prenez pas cette initiative à la légère.

Chapitre 2 : Approches de l'intégration de données

DANS CE CHAPITRE :

- Qu'est-ce que l'intégration de données et comment la mettre en œuvre ?
- Approche traditionnelle (ETL) et approche moderne (ELT)
- Comment l'ELT automatisé résout la problématique d'intégration de données

Principes fondamentaux de l'intégration de données

Le processus d'**intégration de données** se compose des étapes suivantes :

1. Les données sont collectées depuis des flux de capteurs, saisies manuelles ou 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 un entrepôt de données.
3. Les données sont nettoyées et modélisées pour répondre aux besoins d'analyse des diverses entités de l'entreprise.
4. Les données sont utilisées pour alimenter des produits ou solutions de BI.

L'intégration de données peut être réalisée via des processus ad hoc manuels ou programmatiques, à l'aide d'outils logiciels. L'approche ad hoc est imprévisible et non évolutive, tandis que l'approche programmatique nécessite une **pile de données** contenant un ensemble distinct d'outils complémentaires. Ce chapitre traite de ces concepts, ainsi que de l'histoire et de l'avenir de l'intégration de données et des piles de données.

Approches non évolutives de l'intégration de données

De nombreuses entreprises utilisent une approche ad hoc manuelle de l'intégration de données — Dans les faits, 62 % d'entre elles utilisent des tableurs, tels qu'Excel et Google Sheets pour rapprocher et visualiser les données issues de divers fichiers.[2] Ce processus implique de télécharger les fichiers, de modifier ou de nettoyer manuellement les valeurs, de produire des fichiers intermédiaires ou de réaliser des opérations similaires. L'intégration de données ad hoc comporte de nombreux inconvénients ; elle est notamment :

- uniquement adaptée à de très faibles volumes de données ;
- lente ;
- exposée aux erreurs humaines ;
- susceptible de compromettre des informations confidentielles ;
- souvent imprévisible.

Une approche plus viable consiste à entretenir des silos entre les diverses sources de données tout en les reliant via des requêtes « fédérées » qui interrogent directement plusieurs systèmes source et fusionnent les données à la volée. À cette fin, les entreprises peuvent utiliser des moteurs de requêtes SQL, tels que Presto. Toutefois, l'inconvénient de cette approche fédérée est qu'elle implique de nombreux composants mobiles et que ses performances se dégradent avec l'accroissement des volumes de données.

En réalité, une solution évolutive et durable d'analyse de données nécessite une approche systématique et reproductible de l'intégration de données — une pile de données.

Intégration via une pile de données

Une **pile de données** est constituée d'outils et de technologies intégrant et analysant collectivement les données depuis diverses sources. Voici quelques composants d'une pile de données :

1. **Sources de données :**
   a. Applications
   b. Bases de données
   c. Fichiers
   d. Événements numériques
2. **Pipeline de données et connecteurs de données.** Logiciels utilisés pour extraire les données d'une source et les charger dans un entrepôt de données. Cela couvre en majeure partie le processus d'intégration de données.
3. **Entrepôt de données** et/ou **lac de données.** Système de registre conçu pour accueillir d'importants volumes de données de façon permanente. Les entrepôts de

2 zdnet com/article/spreadsheets-still-dominate-business-analytics/

données sont presque toujours orientés colonnes et stockent les données dans une structure relationnelle, tandis que les lacs de données sont des magasins d'objets pouvant contenir à la fois des données structurées et non structurées (brutes).

4. **Modélisation** et/ou **transformations de données.** Il est souvent nécessaire de préparer les données en appliquant une logique métier personnalisée, consistant par exemple à modifier les noms de colonne ou à effectuer des agrégations, afin de les adapter à vos processus d'analyse.

5. **Outil de BI.** Logiciel destiné à synthétiser, à visualiser et à modéliser les données afin d'orienter les décisions métier.

ATTENTION : *Les lacs de données et les entrepôts de données sont conventionnellement utilisés pour stocker différents types de données destinés aux divers cas d'utilisation. En règle générale, les lacs de données contiennent des données brutes, non structurées et sont par conséquent obscurs. Ces données non structurées n'ont pas été nettoyées, normalisées ou transformées avant d'atteindre le système de destination, laissant aux experts des données le soin de les rendre exploitables. Les entrepôts de données contiennent des tables organisées par colonnes et reposent généralement sur des bases de données relationnelles conventionnelles pouvant être interrogées via un système SQL.*

Plus récemment, les lacs de données et les entrepôts de données ont commencé à évoluer de manière convergente. Les lacs de données se sont vus intégrer des transactions ACID (atomicité, cohérence, isolation, durabilité) et des fonctionnalités d'application de schémas visant à « clarifier » les données. Par ailleurs, les entrepôts de données, déjà capables de réaliser des transactions ACID, ont commencé à prendre en charge des outils de science des données et des langages généralement associés aux lacs de données, tels qu'Apache Spark et Python.

Outre cette évolution convergente, les lacs de données sont plus adaptés aux cas d'utilisation pour lesquels la prise en charge de l'apprentissage automatique, de l'intelligence artificielle et d'un écosystème ouvert d'outils de science des données prime sur l'accessibilité. Les utilisateurs de lacs de données sont généralement des experts hautement qualifiés dotés d'une solide expérience de Spark, Python, Pandas et outils similaires pour analyser un large éventail de types de données à grande échelle. Les entrepôts de données sont plus adaptés aux processus d'analyse opérationnelle et d'informatique décisionnelle, pour lesquels les utilisateurs finaux exploitent principalement des systèmes SQL et tableaux de bord BI.

Comment les données se déplacent-elles au sein de la pile

L'unité la plus élémentaire d'un pipeline de données est un élément logiciel appelé **connecteur de données.** Un pipeline de données peut comprendre un ou plusieurs connecteurs, chacun extrayant des données depuis une source — une application,

un moniteur d'événements, un fichier ou une base de données — et appliquant généralement un processus de normalisation et de nettoyage de surface.

Les données sont ensuite acheminées vers un **entrepôt de données**. Les **transformations** peuvent être appliquées avant que les données n'atteignent l'entrepôt ou après leur stockage au sein de ce dernier. C'est ce qui distingue les processus ETL et ELT — sujet que nous aborderons plus en détail par la suite. Dans les deux cas, les transformations peuvent être orchestrées — c'est à dire, organisées en une séquence dont la coordination, la chronologie et les erreurs sont gérées par une logique automatisée. Dans l'idéal, les entrepôts de données font office de système de registre à l'échelle de l'entreprise. N'importe quel type de base de données relationnelle peut être utilisé pour remplir ce rôle ; toutefois, les entrepôts de données sont généralement structurés en colonnes, contrairement aux bases de données transactionnelles ou de production, plus souvent orientées lignes et par conséquent moins efficaces pour traiter les requêtes d'analyse.

Enfin, les données sont analysées à l'aide d'un **outil d'informatique décisionnelle**. Les outils de BI affichent généralement les tendances, pourcentages et autres statistiques sur des tableaux de bord et rapports périodiques.

Chaque composant d'une pile de données peut être hébergé sur site ou dans le cloud. En règle générale, les entreprises utilisent des piles de données sur site. Si le cloud a gagné en popularité, certaines entreprises continuent de mettre en œuvre les composants essentiels de leur infrastructure sur site, afin de répondre aux exigences réglementaires ou à des besoins de performances hautement spécifiques, mais également pour éviter les dépendances externes ou l'enfermement propriétaire.

CONSEIL : *Les considérations techniques distinguant les bases de données orientées lignes de celles orientées colonnes sont en dehors du champ d'application de ce guide ; toutefois, les principes fondamentaux sont les suivants : les bases de données orientées lignes — également appelées bases de données de traitement transactionnel en ligne (OLTP) — sont généralement utilisées pour gérer les transactions dans le secteur de la production. Les bases de données orientées colonnes permettent de mieux gérer les opérations colonnaires propres à l'analyse de données (MIN, MAX, SUM, COUNT, AVG). La familiarisation avec ce type de structure peut vous inciter à faire une simple copie de votre base de données de production à des fins d'analyse. Cette méthode est à proscrire ! Utilisez toujours une base de données ou un entrepôt de données orienté colonnes pour vos opérations d'analyse. Cette approche est bien plus efficace et vous permettra de gagner un temps précieux.*

Problématiques résolues par une pile de données

Lorsqu'elle transmet les données des connecteurs aux entrepôts, la pile de données

doit garantir qu'elles sont centralisées au sein d'un environnement unique et qu'elles restent aussi actuelles et fidèles à la source que possible. Ce processus doit être exécuté en continu, avec une intervention humaine minimale.

Fragmentation

Les données issues des applications, outils et bases de données sont souvent fragmentées. Il existe deux types de fragmentation ; le premier vient du fait que les points de terminaisons d'API et les bases de données opérationnelles ne sont pas conçus pour les requêtes d'analyse, c'est-à-dire que les données qu'ils génèrent n'offrent souvent aucun contexte significatif et ne sont pas organisées pour faciliter les opérations d'analyse. Une modélisation étendue des données est généralement nécessaire pour les interpréter.

Le second type de fragmentation se produit car la plupart des applications, outils et bases de données ne sont pas spécifiquement conçus pour offrir une interopérabilité avec les données d'autres systèmes. Par conséquent, l'établissement du contexte nécessaire en regroupant les registres de plusieurs sources peut entraîner d'importants délais dans le processus de génération de rapports.

Cela se traduit par le phénomène de « données obscures », intervenant lorsqu'une large part des actifs informationnels collectés par une entreprise demeure inexploitée. La centralisation des données au sein d'un même environnement accélère la génération de rapport, tout en permettant aux entreprises de regrouper les registres et d'établir des perspectives cohérentes concernant leurs opérations et clients. Le courtier immobilier Zoopla a par exemple combiné ses données ERP et CRM pour produire un tableau de bord hebdomadaire comprenant plus de 40 KPI distincts.

ÉTUDE DE CAS : DiscoverOrg n'utilise plus sa base de données OLTP pour l'analyse des données

DiscoverOrg est une plateforme de génération de pistes B2B établissant des profils de personnes et d'entreprises afin d'optimiser les campagnes marketing et commerciales. Avant de créer une pile de données, DiscoverOrg récupérait les données d'analyse depuis une copie de sa base de données de production OLTP en excluant celles issues d'applications tierces. Le processus d'interrogation pouvait prendre jusqu'à 36 heures et provoquer un plantage du système.

Suite à l'adoption d'un outil d'intégration de données automatisée, DiscoverOrg a pu combiner les données de production et celles issues de sources tierces au sein d'un même entrepôt de données. Cette approche lui a permis d'économiser les efforts de deux ou trois ingénieurs de données, de générer des rapports en quelques minutes au lieu de plusieurs jours et de développer un algorithme de routage de pistes augmentant le taux moyen d'acquisition de contrats de 80 à 90 %.

Plus récemment, DiscoverOrg a commencé à intégrer des tableaux de bord d'analyse à sa plateforme au profit de ses clients.[3]

Précision

Une imprécision des données peut avoir deux causes ; la première est une mesure ou une consignation incorrecte, en particulier si les données ont été saisies manuellement ou transcrites depuis des supports non numériques.

Les sondages et formulaires sont susceptibles de générer des fautes d'orthographe, des transpositions de caractères et autres erreurs d'ordre administratif. La seconde source d'erreurs, cette fois plus systématique, provient des calculs ou transformations réalisés sur les données brutes. Un jeu de données peut être manipulé de bien des manières, chaque calcul vous éloignant un peu plus des valeurs d'origine. Ainsi, diverses personnes et équipes opérant au sein d'une entreprise peuvent interpréter une même vérité de façon radicalement différente.

Données obsolètes

Les conditions externes évoluent rapidement. Si vous passez des semaines ou des mois à préparer un rapport, vous risquez de prendre des décisions bien mal avisées, car vous travaillerez avec des données obsolètes. Les utilisateurs familiarisés avec les modèles décisionnels tels que la méthode PDCA (Planifier/Développer/Contrôler/Ajuster) ou la boucle OODA (Observer/s'Orienter/Décider/Agir) comprennent l'importance de prendre des décisions avisées plus rapidement que la concurrence. Ces types de modèles s'appliquent à tous les environnements compétitifs et dynamiques, notamment dans les secteurs militaire, des jeux, de l'athlétisme, et bien entendu des affaires.

ÉTUDE DE CAS : Zoopla utilise un pipeline de données pour unifier ses efforts d'intégration des données

Zoopla est un marché immobilier en ligne permettant aux utilisateurs d'acheter, de vendre ou de louer une propriété résidentielle ou commerciale au Royaume-Uni.

Avant que Zoopla n'adopte une pile de données moderne, ses efforts d'analyse étaient dispersés. Ses analystes et ingénieurs ont créé une multitude de scripts ad hoc personnalisés pour extraire et analyser les données. Ces scripts n'étaient pas documentés et souvent écrits dans divers langages. Les analystes utilisaient également des connecteurs de données natifs dans l'outil de BI de Zoopla pour exécuter des interrogations fédérées.

L'équipe BI de Zoopla s'est rendu compte que cette organisation n'était plus viable à mesure que l'entreprise se développait et ajoutait des

3 Consultez l'ensemble de cette étude de cas sur la page fivetran.com/blog/case-studydiscoverorg

sources de données tout en s'efforçant de quantifier sa progression. Suite à l'adoption d'une pile de données moderne, Zoopla a pu combiner les données de ses logiciels ERP et CRM pour produire un tableau de bord actualisé automatiquement sur une base hebdomadaire et comprenant plus de 40 KPI distincts à travers l'entreprise. Ces KPI sont affichés en permanence au sein des bureaux et utilisés par la haute direction comme par les employés les moins qualifiés pour orienter les décisions.[4]

Le coût des opportunités

Les données ne vous servent à rien si vous ne pouvez pas les convertir en perspectives exploitables.

Par le passé, les analystes et ingénieurs n'ont toutefois pas consacré beaucoup de temps à l'analyse des données — au lieu de cela, ils ont investi des efforts considérables dans la création et la gestion de logiciels complexes destinés à les traiter. La science des données est communément associée à la modélisation prédictive de pointe et à l'apprentissage automatique ; néanmoins, les experts passent en moyenne environ 80 % de leur temps à rechercher et à intégrer des données au lieu de les analyser.[5]

ATTENTION : *Le paradoxe de Simpson (Figure 2.0) illustre parfaitement comment les mêmes données, transformées de diverses manières, peuvent aboutir à des conclusions radicalement différentes, voire même totalement opposées.*

En bref, le paradoxe de Simpson décrit le phénomène selon lequel les tendances et schémas paraissent très différents selon le mode de répartition et de combinaison des données.

Un concept similaire est illustré par le quartet d'Anscombe (Figure 2.1) : quatre jeux de données très différents dotés d'une moyenne, d'une variance, d'un coefficient de corrélation et d'un coefficient de détermination identiques. Le paradoxe de Simpson et le quartet d'Anscombe rappellent de façon irréfutable que confiner votre analyse à des statistiques synthétiques élémentaires est au mieux une approche naïve et au pire fortement trompeuse. Vous devez impérativement prendre la peine de visualiser vos données, d'appréhender la manière dont elles sont catégorisées et d'envisager les variables sous-jacentes susceptibles de compliquer l'équation.

L'approche traditionnelle de l'intégration de données : le processus ETL

L'approche traditionnelle de l'intégration de données, appelée processus ETL (Extraire-Transformer-Charger), est prédominante depuis les années 1970. Il s'agit d'un standard

4 Consultez l'ensemble de cette étude de cas sur la page fivetran.com/blog/case-study-zoopla

5 infoworld.com/article/3228245/the-80-20-data-science-dilemma html

sectoriel entre organismes établis, et l'acronyme ETL est communément utilisé pour décrire les activités d'intégration de données dans leur globalité. Le processus ETL a émergé à une époque où les capacités de calcul, de stockage et de bande passante étaient des ressources aussi rares qu'onéreuses. Les défauts techniques de l'ETL, découlant de cette sévère pénurie, semblent toutefois de plus en plus anachroniques à l'ère du cloud.

Figure 2.0
Paradoxe de Simpson

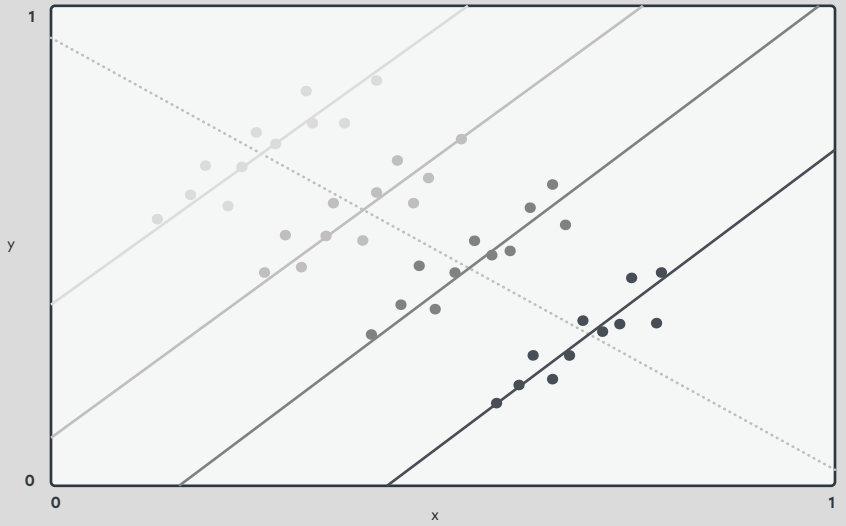
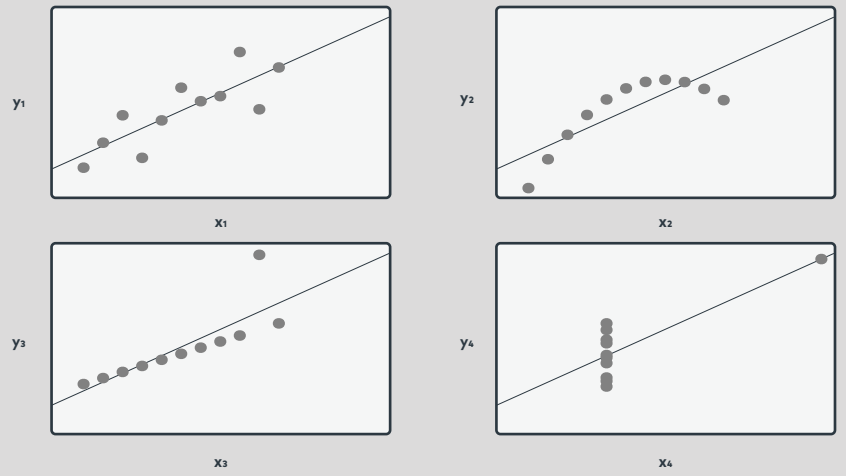


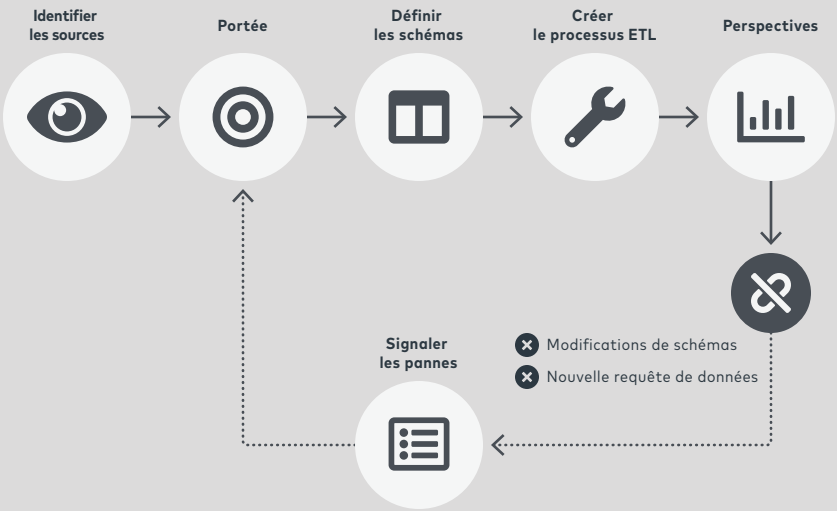
Figure 2.1
Quartet d'Anscombe



Flux de travail ETL

Le flux de travail ETL utilisé par les ingénieurs et analystes pour produire un pipeline ETL ressemble à ce qui suit :

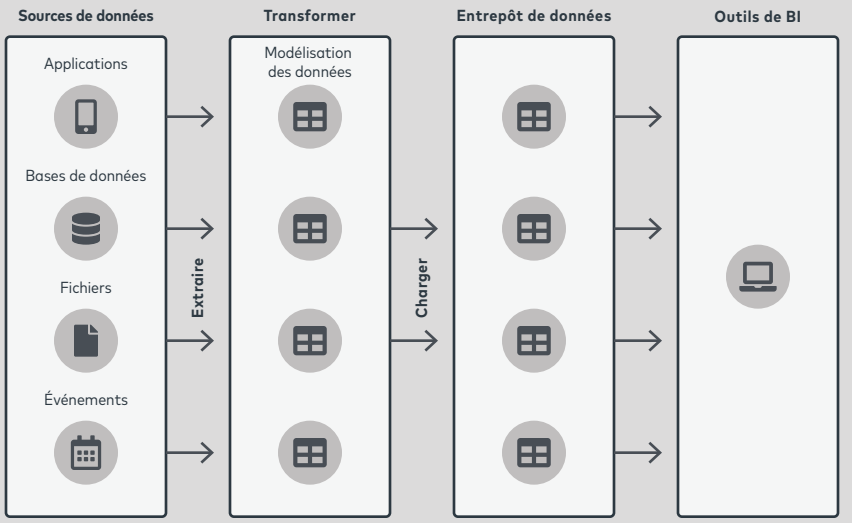
Figure 2.2
Flux de travail d'intégration et d'analyse de données



1. **Identifier les sources** – applications, moniteurs d'événements ou bases de données
2. **Portée** – déterminer les limites et objectifs métier du rapport
3. **Définir les schémas** – modéliser les données et identifier les transformations nécessaires
4. **Créer le processus ETL** – écrire le logiciel en spécifiant les détails des points de terminaison d'API à appeler, comment normaliser les données et les charger dans le système de destination
5. **Perspectives de surface**– générer des rapports assimilables par les décisionnaires clés
6. **Signaler les pannes** – les interruptions privent les utilisateurs finaux de données ponctuelles et provoquent des temps d'arrêt découlant de :
   a. Modifications de schémas en amont
   b. Nouvelles requêtes de données liées à l'évolution des besoins d'analyse
7. **Redimensionner le projet**

Le système ETL exécute les opérations suivantes :

Figure 2.3
Processus ETL



1. **Extraire** – les données sont extraites des connecteurs
2. **Transformer** – les données sont réorganisées en modèles conformément aux besoins des analystes et utilisateurs finaux via une série de transformations
3. **Charger** – les données sont chargées dans un entrepôt de données
4. **Visualiser** – les données sont synthétisées et visualisées dans un outil de BI

L'orchestration et la transformation des données avant leur chargement exposent le processus ETL à une vulnérabilité critique. Les transformations doivent être spécifiquement adaptées aux configurations de la source et de la destination des données. Cela signifie que la modification des schémas de données en amont, ainsi que celle des exigences métier et modèles de données en aval risquent d'endommager le logiciel exécutant les transformations.

Étant donné que l'ETL ne réplique pas directement les données de chaque source dans l'entrepôt de données, il n'existe aucun système de registre exhaustif pour les opérations d'analyse. Une panne survenant à n'importe quelle étape du processus rendra les données inaccessibles aux analystes et nécessitera des efforts de réparation de la part des ingénieurs.

Limitations du processus ETL

Globalement, le processus ETL traditionnel présente trois inconvénients majeurs et connexes :

1. **Complexité.** Les pipelines de données exécutent du code personnalisé répondant aux besoins de chaque transformation. Cela implique que l'équipe d'ingénierie des données développe des compétences hautement spécialisées et parfois non transférables pour gérer sa base de code.
2. **Instabilité.** Pour les raisons mentionnées ci-dessus, un mélange d'instabilité et de complexité rend les ajustements rapides coûteux, voire impossibles. Certaines parties de la base de code peuvent devenir inopérationnelles de façon inopinée, et les nouveaux besoins métier ou cas d'utilisation nécessitent sa révision en profondeur.
3. **Inaccessibilité.** Et surtout, l'ETL est un processus quasi inaccessible aux sociétés de faible envergure sans ingénieurs de données dédiés. L'ETL sur site impose des coûts infrastructurels supplémentaires, ce qui peut forcer les petites entreprises à échantillonner les données ou à générer manuellement des rapports ad hoc.

L'émergence de la technologie cloud

Même un observateur distant des tendances technologiques sait que les capacités de calcul, de stockage et de bande passante sont désormais des ressources omniprésentes et bon marché. Avec le développement de l'informatique, leur coût a chuté au fil des ans (Figure 2.4).

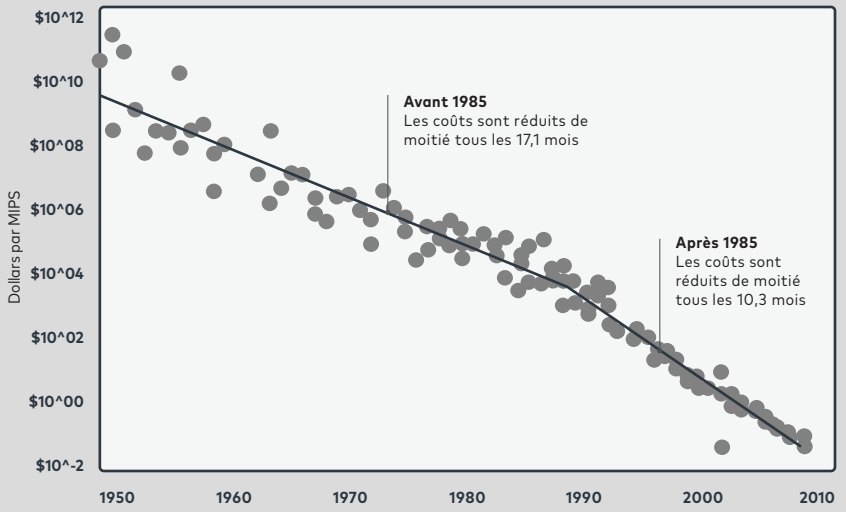
Par ailleurs, sur une période d'environ 35 ans, le coût du gigaoctet est passé de quasiment 1 million $ à quelques cents (Figure 2.5).

L'une des conséquences de cette réduction de coût radicale est la capacité des entrepôts de données à stocker des volumes bien plus vastes. Les entreprises n'ont plus besoin de pré-agréger les données et éliminent au passage une part significative des données source, ce qui permet aux experts de réaliser des analyses plus approfondies et exhaustives que jamais.

Si le réseau mondial n'existe que depuis 1991, le coût du trafic sur Internet a également considérablement diminué. En moins de vingt ans, il a chuté d'environ 1 200 $ par Mbit/s à quelques cents (Figure 2.6).

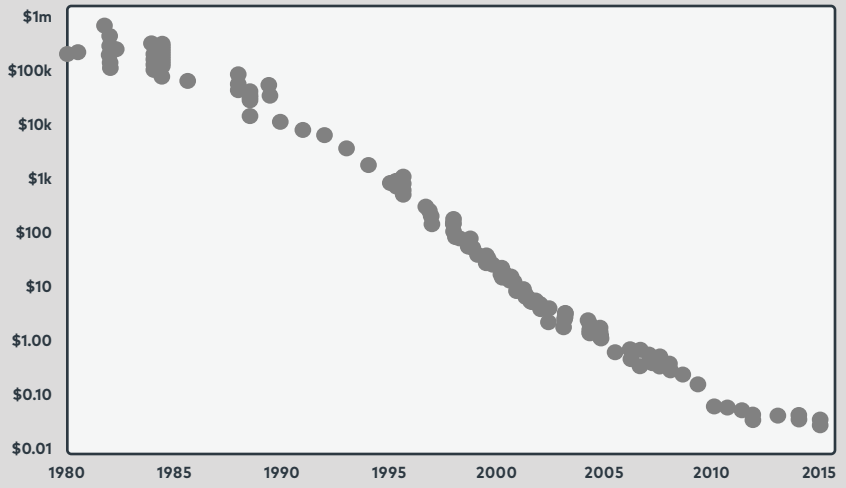
La convergence de ces trois tendances de réduction des coûts a abouti au cloud — c'est-à-dire, à l'utilisation de ressources informatiques Web distantes et décentralisées. La technologie cloud, à son tour, a engendré une multitude d'applications et de services natifs.

Figure 2.4
Coût de l'informatique



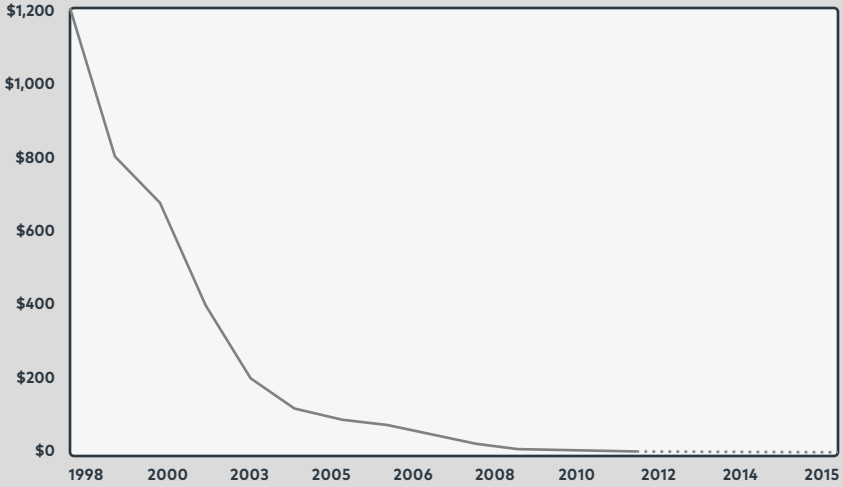
Source : https://frc.ri.cmu.edu/~hpm/book97/ch3/processor.list.txt

Figure 2.5
Coût du disque dur par gigaoctet



Source : mkomo.com

Figure 2.6
Prix du trafic Internet (bande passante)



Source : dr1peering.net

CONSEIL : *Les applications et services cloud natifs englobent tous les types d'activités professionnelles, notamment la gestion des relations client, la facturation et le paiement, l'e-commerce, le marketing par e-mail, l'administration des avantages, la gestion de projets et le service client. Il y a de fortes chances pour que votre entreprise utilise déjà plusieurs de ces services.*

L'un des principaux avantages offerts par le cloud est que les analystes et utilisateurs finaux des données ne sont plus confinés aux infrastructures physiques. Au lieu de cela, ils peuvent héberger des services sur le Web, ce qui simplifie sensiblement les problématiques de mise à l'échelle et d'accessibilité. Les entreprises peuvent ajouter ou supprimer des ressources de traitement et de stockage à la volée, et les utilisateurs peuvent accéder à des tableaux de bord et rapports via n'importe quel appareil connecté.

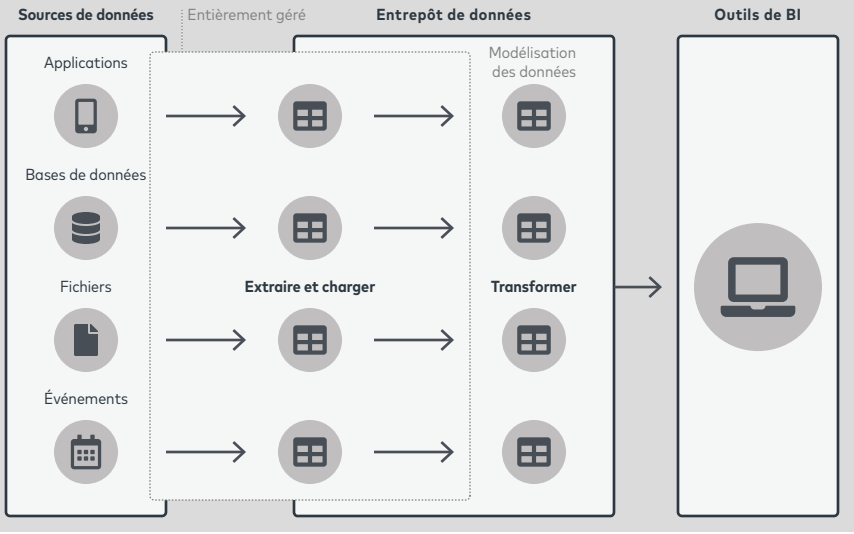
ATTENTION : *Ne surestimons pas les capacités de l'Internet moderne. Nous ne disposons pas (encore) de la technologie nécessaire pour charger et télécharger instantanément des téraoctets de données à travers le monde. À très grande échelle, il peut être plus rapide d'expédier physiquement un conteneur rempli de disques durs que d'envoyer les données via Internet. Il est toujours essentiel de compresser les données et de minimiser le volume transféré de la source à la destination.*

L'approche moderne de l'intégration de données : le processus ELT

Les tendances ayant favorisé le développement du cloud — la réduction drastique des coûts en matière de traitement, de stockage et de bande passante — ont également permis aux entreprises de contourner les problèmes posés par l'ETL. Plus particulièrement, les données peuvent être diffusées via des flux et chargées avant d'être transformées. Cette séquence révisée, appelée ELT (Extraire-Charger-Transformer), est le successeur moderne de l'ETL (Figure 2.7).

Le composant appelé « pile de données moderne » est basé sur l'ELT et remplace les technologies sur site par des solutions SaaS cloud native. Ce type de configuration facilite l'automatisation, la collaboration et l'évolutivité tout en éliminant une part substantielle des coûts liés à l'implémentation d'une pile de données sur site. Correctement mise en œuvre, la pile de données moderne offre une intégration de données et une accessibilité continues à l'échelle de l'entreprise, tout en limitant au minimum les interventions manuelles et l'utilisation de code personnalisé.

Figure 2.7
Processus ELT



La permutation les étapes de chargement et de transformation résout chacun des trois inconvénients majeurs de l'ETL :

1. **Complexité.** Le pipeline est simplifié — l'envoi initial des schémas standard vers

l'entrepôt de données, sans transformations personnalisées, permet de transférer une part importante des tâches liées au pipeline en aval, afin qu'elles soient traitées par les analystes et non par les ingénieurs de données.

2. **Instabilité.** Le pipeline est plus résilient et plus fiable — les transformations étant appliquées une fois les données stockées dans l'entrepôt, les pannes provoquées par les modifications du système source affectent principalement la couche analytique et les analystes peuvent généralement résoudre ces problèmes sans l'aide d'ingénieurs des données.

3. **Accessibilité.** Le pipeline est plus accessible, car sa gestion nécessite moins de ressources. Étant donné qu'il est considérablement simplifié et intrinsèquement plus résilient, les tiers peuvent créer et gérer un outil standardisé pour plusieurs clients, ainsi que des produits dérivés visant à optimiser les initiatives d'analyse. En essence, l'achat d'un outil standardisé externalise et automatise les phases d'extraction et de chargement.

Les transformations d'entrepôt intégrées permettent de créer des tables dérivées appelées « vues » sans altérer les données de la source. Les entreprises peuvent alors créer un système de registre insensible à l'évolution des besoins métier ou aux modifications de schémas en amont. Les mêmes données peuvent être appliquées à plusieurs cas d'utilisation.

L'ELT réduit également la charge de travail des ingénieurs. Une fois les données stockées dans l'entrepôt, les analystes peuvent utiliser un système SQL pour exécuter les transformations souhaitées. Des transformations complexes devant être soigneusement orchestrées et planifiées peuvent toujours être nécessaires, mais les interruptions et les pannes n'entravent plus l'ensemble du pipeline de données et n'impliquent plus d'efforts significatifs de la part des ingénieurs.

Une approche optimisée : le processus ELT automatisé

La nature simplifiée et basée sur le cloud d'une pile de données ELT est propice à l'automatisation et à l'externalisation.

Les activités spécifiques liées à l'approche ELT automatisée comprennent la détection et la réplication des modifications, le nettoyage de surface et la normalisation des données, ainsi que la mise à jour et la création des tables. Ces activités nécessitent une profonde connaissance des sources de données, un grand savoir-faire en matière de modélisation et d'analyse des données, ainsi que les compétences d'ingénierie requises pour créer des systèmes logiciels robustes. Sans outil d'intégration de données automatisée, votre équipe devra elle-même réaliser ces tâches et développer les fonctionnalités requises.

En revanche, l'automatisation et l'externalisation du processus ELT vous permettent de bénéficier du savoir-faire de tiers comprenant toutes les spécificités des sources de données sous-jacentes — et ayant testé leurs connecteurs sur un panel de cas marginaux bien plus vaste que vous ne pourrez sans doute le faire. À l'instar de nombreuses formes d'automatisation, le processus ELT automatisé offre des économies de temps, d'efforts et d'argent. Votre équipe de données ou de BI doit se concentrer sur le développement de perspectives exploitables, et non sur des tâches de routine en amont traitant des problèmes ayant déjà été identifiés et résolus.

Les ingénieurs des données peuvent tirer parti des gains de temps offerts par l'ELT automatisé pour concentrer leurs efforts sur les problématiques affectant les clients externes ou les investir dans des activités à plus forte valeur, telles que l'apprentissage automatique et l'intelligence artificielle. Il est plus pertinent d'envisager L'ELT automatisé comme un amplificateur de performances, plutôt qu'une alternative aux talents humains.

L'accessibilité totale offerte par l'ELT automatisé est illustrée par le flux de travail typique de l'intégration de données automatisée :

1. **Inscription** – Activez votre compte
2. **Sélection** – Choisissez vos sources et votre entrepôt de données
3. **Authentification** – Activez vos connexions à l'aide des identifiants existants
4. **Automatisation** – Laissez le système gérer la synchronisation historique et les modifications en continu

Si une synchronisation historique complète peut prendre plusieurs heures ou jours selon le volume de données hébergé par la source, les phases nécessitant réellement une intervention humaine ne prennent que quelques minutes.

Autre avantage d'un outil d'intégration de données automatisé, chaque entreprise utilisant une source de données spécifique doit résoudre exactement le même problème, ce qui permet aux fournisseurs d'offrir à chacun de leurs clients une solution unique et standardisée dotée de schémas identiques. Ces schémas standardisés permettent la création de produits dérivés optimisant les opérations d'analyse. Les utilisateurs d'un même outil d'intégration auront accès aux mêmes transformations SQL, produits d'analyse intégrés et modules de solution BI. L'ELT offre aux analystes les avantages et la portée d'un écosystème de composants modulaires et interchangeables.

Le processus ELT automatisé offre un dernier avantage : il vous permet d'externaliser la cybersécurité et la conformité réglementaire. Le développement des stratégies, procédures et technologies empêchant tout accès malveillant ou illicite à vos données nécessite un grand savoir-faire, et il est nettement préférable de confier

cette tâche à un tiers digne de confiance qu'entreprendre la création de votre propre solution.

Le fait de délivrer des fonctionnalités d'automatisation et de libre-service aux ingénieurs, analystes et utilisateurs finaux augmente toutefois l'importance de la gouvernance des données. L'accessibilité et la transparence peuvent être de précieux atouts, mais ils doivent être gérés via une politique stricte en matière d'audit, de documentation et d'affectation des autorisations. Au fur et à mesure du développement d'une entreprise, les analystes, qui organisent généralement les données en tableaux de bord et rapports, peuvent se voir affecter des fonctions de gouvernance des données, alors que les utilisateurs finaux assument de plus en plus eux-mêmes les tâches de génération de rapports et de tableaux de bord.

ÉTUDE DE CAS : DocuSign triple ses sources de données grâce à l'intégration de données automatisée

Leader mondial en technologie de signature numérique, DocuSign aide les personnes et les entreprises à préparer, signer, amender et gérer automatiquement les contrats.

Auparavant, DocuSign utilisait SQL Server comme entrepôt de données et disposait de six sources de données gérées par un ingénieur. Le développement de ces connecteurs artisanaux prenait trois à six mois et leur maintenance jusqu'à 20 heures d'ingénierie par semaine. Au fur et à mesure du développement de l'entreprise, la charge de travail est devenue ingérable, en particulier car les compétences des ingénieurs étaient requises pour les projets critiques et que les équipes commerciales devaient modéliser et cataloguer les données collectées depuis les applications.

En adoptant une solution d'intégration de données automatisée et un entrepôt cloud plus élastique, DocuSign a pu économiser ses 20 heures d'ingénierie et tripler le nombre de sources de données, passant de 6 à 18. L'accroissement soudain de l'infrastructure, les gains de temps et les économies de main-d'œuvre se sont accompagnés d'une autre avancée essentielle — tous les collaborateurs de l'entreprise utilisent désormais plus de 100 tableaux de bord actifs dans leur outil de BI.[6]

N'OUBLIEZ PAS : *La technologie cloud a généré une mine de données exploitables, mais également des outils adaptés pour les manipuler. L'ELT élimine de nombreux inconvénients de l'ETL, en rendant les données et l'analyse plus accessibles et évolutives que jamais.*

6 Consultez l'ensemble de cette étude de cas sur la page fivetran.com/blog/case-study-docusign

Chapitre 3 : Pourquoi vous ne devez pas créer votre propre pipeline de données

DANS CE CHAPITRE :

- Comment estimer les coûts monétaires et non monétaires liés au développement de votre propre pipeline de données.
- Convaincre votre entreprise d'adopter une solution commerciale

Considérations clés

Si la pile de données moderne simplifie radicalement l'intégration de données, est-il pertinent pour votre entreprise de créer son propre pipeline ELT, même dans le cloud ?

ATTENTION : *N'oubliez pas que l'intégration de données manuelle n'est pas évolutive et que le processus ELT a été supplanté par les tendances technologiques prédominantes, comme expliqué au chapitre 2. Ce chapitre traite principalement de la création d'un pipeline ELT sur mesure, bien que les arguments suivants s'appliquent également à une entreprise tentant de créer un flux de travail ELT personnalisé.*

Délais et coûts

Comme mentionné au chapitre 2, un expert des données consacre en moyenne 80 % de son temps à la mise en œuvre de pipelines de données — une tâche pour laquelle les aptitudes, l'intérêt et la formation de la plupart d'entre eux sont limités (Figures 3.0 et 3.1). L'argument le plus évident contre la création d'un pipeline ELT

personnalisé est son coût de mise en œuvre et de gestion en termes d'argent, de temps, de motivation et de perte d'opportunités.

Supposons que votre entreprise nécessite cinq connecteurs supplémentaires pour la gestion des relations client, la génération de tickets de support client, l'automatisation publicitaire, la gestion de projet et la facturation des abonnements.

La création de chacun des cinq connecteurs prend environ cinq semaines à un ingénieur, ou cinq semaines-homme (sh) :

```
(5 connecteurs) x (5 sh)
```

Chaque connecteur nécessitera vraisemblablement une semaine de maintenance trimestrielle, ajoutant jusqu'à quatre semaines par an :

```
(5 connecteurs) x (5 sh + 4 sh)
(5 connecteurs) x (9 sh) = 45 sh
```

Cela représente 45 semaines sur 52 par an. Ajoutez à cela un arrêt de travail ou une politique de congé maladie légèrement permissive, et nous arrivons à une année entière de travail pour un ingénieur informatique, à hauteur de 120 000 $, sans même aborder les avantages sociaux (Figure 3.2).

Au cours des années suivantes, votre ingénieur continuera à actualiser les connecteurs chaque trimestre (quatre semaines) et à traiter les bogues et cas marginaux émergeant (une semaine) pour un total de cinq semaines-homme par connecteur.

```
(5 connecteurs) x (5 sh) = 25 sh
```

Cela représente 25 semaines annuelles sur 52 dédiées à la maintenance continue. Estimons grossièrement ce coût à la moitié du salaire annuel d'un ingénieur, soit 60 000 $.

Le coût d'achat ou d'externalisation des cinq connecteurs sera vraisemblablement bien inférieur aux chiffres annoncés ci-dessus. Bien entendu, ces coûts augmenteront proportionnellement au nombre de sources de données que vous utilisez.

Figure 3.0
Sur quelles tâches les experts de données passent-ils le plus clair de leur temps ?

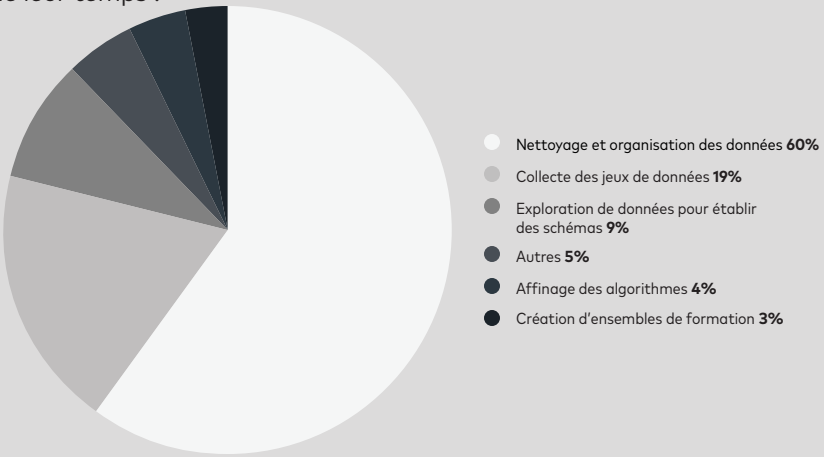
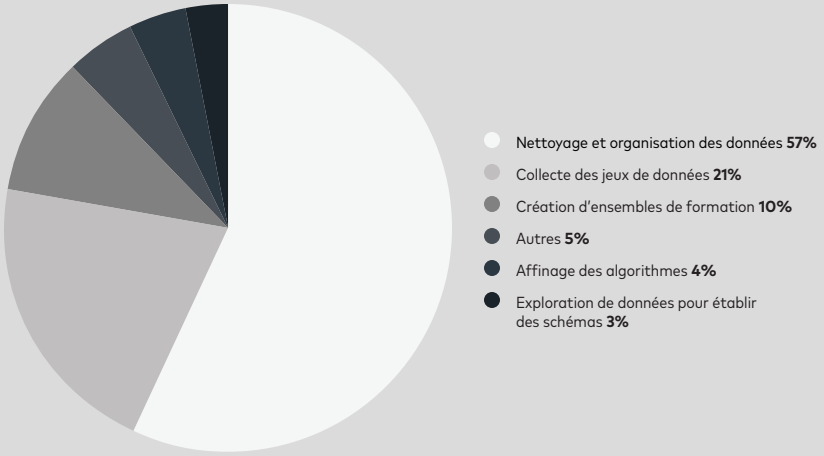
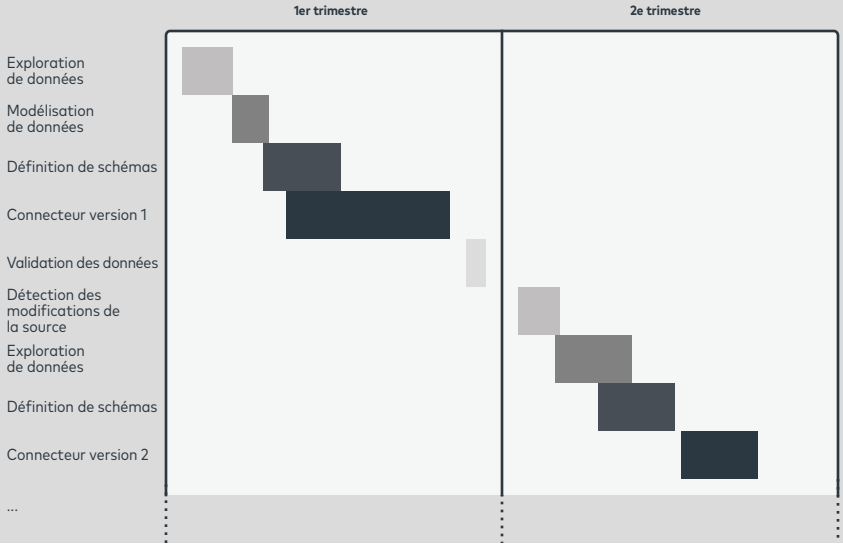


Figure 3.1
Quelle est la part la plus ingrate de la science des données ?



Source: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

Figure 3.2
Quelle charge de travail nécessite la création de votre propre connecteur ?



L'exemple de diagramme de Gantt ci-dessus démontre la nature cyclique et récurrente du travail d'ingénierie, même dans une infrastructure ELT, car les schémas en amont continuent d'évoluer.

Motivation

Si vous souhaitez préserver la motivation de vos analystes, ingénieurs et responsables, envisagez les problèmes suivants liés à la création de vos propres connecteurs ou à la génération manuelle de rapports :

1. Détournement des autres fonctions liées à l'ingénierie des données, la science des données ou l'analyse — problématique très courante parmi les nouveaux experts des données dans les sociétés souffrant de pénuries de personnel et pouvant entraîner une rotation d'effectifs.
2. Frustration et épuisement face à la complexité de la gestion de l'intégrité des données, en particulier pour le personnel sans formation adéquate.
3. Temps d'arrêt provoqués par l'accroissement de la complexité dû à l'ajout inévitable de sources de donnée supplémentaires.
4. Décisions mal avisées provoquées par des décalages entre les requêtes BI et la mise à disposition des perspectives exploitables — il se peut que les perspectives soient obsolètes dès leur disponibilité.

Pour la plupart des professionnels des données, la maintenance des bases de données est une corvée, et non une aspiration.

Courbes d'apprentissage

L'estimation de cinq semaines citée ci-dessus s'applique aux API relativement simples. Certaines ne sont pas aussi maniables, car elles ne tiennent pas compte des meilleures pratiques, sont mal documentées ou simplement très complexes.

Les données issues d'outils de planification des ressources d'entreprise (ERP) peuvent par exemple englober toutes les activités métier imaginables, représentées par des dizaines, voire des centaines de tables dotées d'associations complexes. Le développement d'un logiciel éprouvé pour gérer une telle source de données peut nécessiter de nombreuses itérations, et multiplier les coûts énoncés ci-dessus.

Complexité croissante

Il est peu vraisemblable que les besoins en données de votre entreprise s'arrêtent à cinq connecteurs. Comme mentionné précédemment, les entreprises utilisent en moyenne plus de 100 applications[7], et ce chiffre est susceptible de croître. Il est difficile de justifier l'accroissement des obligations de votre équipe quand vous pouvez externaliser l'ingénierie de votre pipeline à moindre coût.

Standardisation

Le dernier argument en défaveur d'un pipeline de données personnalisé est que les connecteurs développés par des tiers sont robustes, car ils ont été testés sur des dizaines de cas marginaux auprès d'une multitude de clients. Ces connecteurs produisent des jeux de données standardisés avec des schémas normalisés (Figure 3.3).

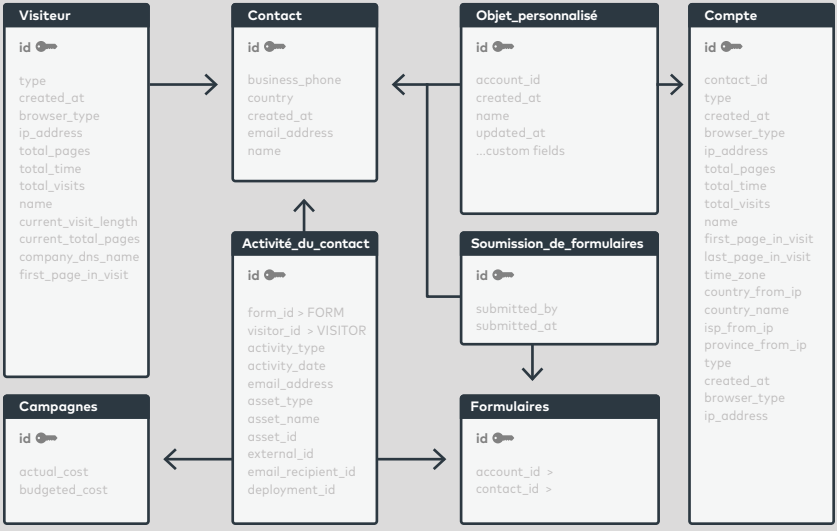
En pratique, il y a peu de chances que vous utilisiez des tables normalisées pour alimenter directement vos tableaux de bord ; vous les transformerez plutôt en modèles mieux adaptés aux utilisateurs finaux.

En principe, cette approche permet également à toutes les entreprises utilisant les mêmes connecteurs de tirer parti des mêmes transformations, car les données sont structurées de façon identique. Les schémas ou modèles prêts à l'emploi de ce type peuvent être écrits en SQL ou dans des langages spécifiques à la plateforme BI, tels que LookML.

En règle générale, la standardisation génère des économies d'échelle, car le fournisseur finit par assimiler toutes les spécificités des sources de données sous-jacentes et en partage les avantages avec ses clients.

7 https://www.wsj com/articles/employees-are-accessing-more-and-more-businessapps-study-finds-11549580017

Figure 3.3
Schéma standardisé



CONSEIL : *Il existe autant de méthodes de création d'un schéma depuis un jeu de données que d'opinions et de cas d'utilisation. L'une des approches pour conserver un standard reproductible et simple à comprendre consiste à normaliser les processus. L'analyse approfondie des formes de normalisation n'est pas l'objet de ce guide, mais le principe fondamental est que la normalisation crée une structure relationnelle qui élimine la redondance et les associations incohérentes entre les éléments de données.*

N'OUBLIEZ PAS : *Si le concept d'extraction et de chargement de données semble simple, les pipelines de données sont des composants technologiques hautement sophistiqués dont le développement nécessite des ressources considérables en termes de compétences, de main-d'œuvre et de spécialisation. Il est à bien des égards préférable d'acheter une solution que de tenter de la créer et de la gérer en interne.*

Avantages liés à l'achat d'une solution d'intégration de données

En proposant l'adoption d'une solution de pipeline de données au sein de votre entreprise,

il se peut que vous soyez confronté à une certaine réticence. Les ingénieurs de données peuvent se soucier de l'automatisation d'une partie de leurs tâches et les membres de la direction, qui n'appréhendent pas directement les tenants et aboutissants de cette problématique, peuvent ne pas en percevoir immédiatement les avantages.

ÉTUDE DE CAS : Papier double ses sources de données et développe un modèle d'attribution client grâce à l'ELT

Papier est une société de design et de personnalisation vendant des articles de papeterie, des invitations, des cartes et des albums photo. Fondée en 2015, la société native du Royaume-Uni a prospéré et développe actuellement ses opérations et sa gamme de produits. La complexité de ses processus d'intégration de données s'est développée avec son activité.

Papier centralisait initialement ses données transactionnelles, publicitaires et ses flux de clics à l'aide de scripts ETL et d'outils développés en interne. Cette approche a rapidement trouvé ses limites en termes de charge de travail et les efforts de correction des imprécisions et incohérences nécessitaient souvent des resynchronisations.

Après avoir adopté des services de pipeline de données ELT, Papier a pu automatiser son flux de travail d'ingestion de données et doubler le nombre de sources utilisées pour l'analyse. Les synchronisations exécutées auparavant une fois par jour le sont désormais toutes les heures. Et surtout, en centralisant ses données et en doublant le nombre de sources, Papier a pu développer un modèle d'attribution client plus robuste, identifiant les publicités et autres activités marketing générant le plus de chiffre d'affaires.[8]

Convaincre les ingénieurs

Les ingénieurs encouragent parfois la personnalisation, la configurabilité granulaire et le contrôle de l'accessibilité. Les arguments suivants peuvent vous aider à les rallier à votre cause :

En quoi la création de pipelines est-elle néfaste

1. La complexité et les délais étendus génèrent des goulets d'étranglement dommageables.
2. Si elles ne sont pas en mesure de prendre rapidement des décisions avisées, les entreprises ne peuvent s'adapter et rester compétitives.
3. Les tâches de traitement des données de routine sont ennuyeuses et démoralisantes.

8 Consultez l'ensemble de cette étude de cas sur la page fivetran.com/blog/case-studypapier

Comment l'externalisation peut vous aider

1. Elle élimine les goulets d'étranglement.
2. Les fournisseurs externes spécialisés dans les connecteurs de données sont sensiblement plus qualifiés et expérimentés pour résoudre ces problèmes. Ils ont sans doute traité bien plus de cas marginaux que vous et seront donc plus efficaces.
3. Le processus ELT est un multiplicateur de forces. Il vous permet de gérer un éventail de connecteurs bien plus vaste avec beaucoup moins d'efforts, facilitant ainsi la satisfaction des attentes de votre entreprise.
4. Le processus ELT entièrement géré est incroyablement intuitif et ne nécessite quasiment aucun effort de formation ou d'enseignement. Vous pourrez ainsi consacrer votre temps à des activités autres que le développement de compétences hautement spécialisées sur des outils.
5. Il n'est pas nécessaire d'engager et de gérer davantage de personnel en ingénierie des données.
6. Vous pourrez consacrer du temps aux activités hautement stratégiques ; par exemple :
    a. Outils et infrastructure personnalisés pour les analystes
    b. Infrastructure prenant en charge l'IA et l'apprentissage automatique
    c. Nouveaux produits logiciels

La meilleure approche pour relever ce défi consiste à privilégier les tâches sur lesquelles les ingénieurs souhaitent se concentrer. En pratique, très peu d'entre eux ont envie de passer le plus clair de leur temps à développer des connecteurs de données. Ils préfèrent les projets à forte valeur aux tâches de développement et de gestion d'infrastructure.

Convaincre votre supérieur

Il peut être nécessaire de convaincre les membres de la direction - qui n'appréhendent pas directement les tenants et aboutissants en matière d'ingénierie des données - du fait que la valeur définitive d'un nouvel outil justifie son prix élevé et les remaniements du personnel susceptibles d'en découler.

Mettre en avant la réussite des autres entreprises

Heureusement, l'expérience des autres entreprises offre une mine d'informations concrètes. Il peut être préférable de commencer par discuter des avantages d'une

solution BI optimisée et des perspectives exploitables avant d'aborder la mise en œuvre de ces améliorations.

Débattre des contraintes courantes et des avantages clés

Les contraintes courantes sont les suivantes :

1. La génération manuelle de rapports est très chronophage et à peine meilleure que la navigation à vue.
2. Diverses unités commerciales utilisent différents silos et échangent difficilement les informations pertinentes.
3. Les outils internes deviennent ingérables face à l'ajout de sources et à l'accroissement du volume de données, et les exigences de performances deviennent plus strictes.
4. Les bases de données héritées et les entrepôts sur site atteignent leurs limites de performances et d'exploitabilité.
5. Les ingénieurs ont mieux à faire que gérer des bases de données.

Avantages courants :

1. Gain de temps – réduction considérable des délais entre chaque rapport.
2. Gain de données – extension massive de la disponibilité et de l'actualisation des données.
3. Gain de qualité – les données sont plus exhaustives et ponctuelles.
4. Gain culturel – l'accès aux données et les décisions axées sur les données sont démocratisés à l'échelle de l'entreprise.
5. Économie de main-d'œuvre – réduction du temps consacré aux outils d'intégration personnalisés et à la maintenance des bases de données ; les analystes n'ont pas à organiser les rapports manuellement.
6. Nouvelles perspectives et nouveaux produits – libérer les ressources d'ingénierie des tâches d'intégration de données se traduit par une augmentation de la bande passante pour explorer des opportunités et développer des produits.

N'oubliez pas que les membres de la direction opèrent souvent en aval du processus d'intégration de données (sauf dans le cas des DSI) ; il est donc essentiel de mettre en avant les avantages d'une solution de BI performante avant d'aborder les aspects techniques liés à l'entreposage et à l'intégration des données.

ÉTUDE DE CAS : MVF double son revenu mensuel préservé grâce à une pile de données moderne

MVF est une plateforme de référencement client fournissant des pistes de ventes avec un modèle de paiement unitaire. Avant la mise en œuvre d'une pile de données moderne, MVF ne disposait d'aucune source unique et fiable, et utilisait des bases et piles de données ad hoc distinctes réparties à travers les équipes. La génération de rapports prenait deux à trois semaines.

Grâce à Fivetran, MVF a pu centraliser ses données et identifier rapidement les pistes invendues. Ces avantages ont permis à l'entreprise de doubler son revenu mensuel, passant de 300 000 à 700 000 £. Les rapports préparés auparavant en deux à trois semaines sont désormais générés automatiquement en continu.

En outre, les ingénieurs de l'entreprise ont économisé quatre à huit semaines initialement consacrées au développement de connecteurs, ainsi que les coûts liés aux processus de maintenance et de débogage continus. L'automatisation de l'intégration de données a également permis à MVF d'ajouter huit sources de données supplémentaires. Les ingénieurs, quant à eux, travaillent à présent sur des projets plus stratégiques à haute valeur.[9]

'OUBLIEZ PAS : *La prospérité à long terme de votre entreprise dépend de sa capacité à suivre la courbe technologique. La culture des données est devenue une composante critique pour rester compétitif sur le marché actuel ; aussi, les obstructions de la direction et les réticences des ingénieurs peuvent en définitive mettre en péril l'ensemble de l'entreprise.*

9 Consultez l'ensemble de cette étude de cas sur la page fivetran.com/blog/case-study-mvf

Chapitre 4 : Considérations métier pour choisir un outil d'intégration de données

DANS CE CHAPITRE :

- Modèle de tarification
- Adéquation aux compétences de votre équipe et à vos plans futurs
- Garantir l'évolutivité

Les économies offertes par une solution d'intégration des données en termes de temps, de coûts et de main-d'œuvre dépendent de la taille et de la maturité de votre entreprise, ainsi que des caractéristiques spécifiques au fournisseur de pipeline.

À très petite échelle, il est probable que votre entreprise ne nécessite pas de pipeline de données, en particulier si vous opérez dans une start-up récente qui n'utilise qu'une ou deux sources de données ou si vous réalisez uniquement des analyses qualitatives à la recherche du segment de marché adéquat. À l'inverse, votre entreprise peut exploiter une niche avec des exigences très strictes en matière de performances, de sécurité ou de conformité réglementaire. Certaines applications de science des données peuvent être extrêmement sensibles à la latence, parfois même à quelques nanosecondes près.

En excluant les deux scénarios ci-dessus, votre entreprise peine probablement à assumer les coûts d'ingénierie liés à la création et à la gestion des connecteurs de données, ou fait face à d'importants délais dus à la maintenance des connecteurs et à la génération manuelle de rapports. Si c'est le cas, il peut être pertinent d'envisager l'acquisition d'une solution d'intégration de données.

Modèles de tarification et coûts

Familiarisez-vous avec les structures de tarification des outils que vous évaluez. Voici quelques modèles de prix courants :

- **Forfait fixe.** Un abonnement peut avoir un prix fixe plus élevé, mais offre des coûts prévisibles.
- **Tarification au volume** de données, en gigaoctets ou lignes. Un modèle de tarification basé sur le volume peut être très avantageux si vous traitez actuellement une quantité de données très modeste, mais que vous souhaitez tester un nouvel outil sur une période étendue, ou prévoyez de transférer progressivement votre flux de travail vers le nouveau système.
- **Tarification par poste et tarif unique** à l'échelle de l'entreprise**.** Les modèles de tarification par poste reviennent moins cher si vous disposez d'effectifs restreints, mais impliquent plus de contraintes administratives. Le modèle de tarif unique à l'échelle de l'entreprise peut être plus simple et moins cher pour les sociétés avec un personnel plus conséquent.

Il se peut également que vous rencontriez des combinaisons de modèles de tarification. Un service peut avoir un abonnement de « plateforme » à prix fixe, puis des frais supplémentaires pour chaque connecteur de données. Les taux liés aux volumes sont susceptibles de varier selon les connecteurs et les fournisseurs peuvent proposer un modèle « freemium » jusqu'à un certain volume de données, ou offrant un ensemble de fonctionnalités restreint. En d'autres termes, votre consommation peut varier.

Adéquation aux compétences de votre équipe et à vos plans futurs

Parmi les autres facteurs à prendre en compte figure le compromis entre simplicité d'utilisation, configurabilité et adéquation avec la gamme de compétences de votre équipe.

Les utilisateurs sans compétences techniques ne seront vraisemblablement pas familiarisés avec les systèmes SQL, mais pourront certainement utiliser un outil de BI. Les analystes connaissent généralement les langages SQL, statistiques et éventuellement de scriptage (ex. : Python). Les experts des données peuvent disposer de compétences techniques plus approfondies, notamment d'une formation statistique plus avancée, et maîtriser des langages supplémentaires, tels que Java, ainsi que des technologies de « mégadonnées », telles qu'Hadoop ou Spark. Les ingénieurs, quant à eux, maîtrisent généralement un éventail de

langages informatiques de haut et bas niveau, ainsi qu'une variété de plateformes technologiques.

Différents outils d'intégration des données impliquent divers niveaux de complexité et d'accessibilité. Certains reposent principalement sur le scriptage personnalisé et n'offrent qu'une structure élémentaire pour soutenir le développement de votre pipeline de données. D'autres délivrent des interfaces graphiques permettant aux utilisateurs non spécialisés d'orchestrer les opérations de réplication et detransformation des données, mais ont deux inconvénients majeurs : une courbe d'apprentissage élevée et hautement spécifique à la plateforme et la génération automatique de code spaghetti. D'autres encore combinent réplication des données entièrement automatisée et transformations SQL avec contrôle de version.

Le compromis se résume à équilibrer accessibilité et configurabilité. Si votre objectif est de promouvoir la culture des données à l'échelle de l'entreprise, il peut être pertinent de rechercher un outil offrant une simplicité d'utilisation optimale et un vaste champ d'applicabilité aux divers cas d'utilisation.

Pour les cas plus spécialisés, des outils moins accessibles, mais plus performants et configurables, optimisés pour des segments spécifiques seront plus adaptés.

Enfermement propriétaire et évolution des besoins

Avant de vous engager, assurez-vous que l'outil que vous avez choisi répondra à vos futurs besoins. Posez les questions suivantes :

- L'outil dispose-t-il des connecteurs que vous utilisez actuellement ou que vous prévoyez d'utiliser ?
- Est-il possible d'ajouter facilement de nouvelles fonctionnalités ou des connecteurs de données à votre compte selon les besoins ?
- Les connecteurs sont-ils mis à jour de façon cohérente pour s'adapter aux modifications d'API en amont, et ces mises à jour sont-elles accompagnées de journaux spécifiant les modifications ?
- De nouveaux connecteurs sont-ils régulièrement ajoutés à l'outil ?
- L'équipe de support est-elle réactive et en mesure de suivre l'évolution des produits comme de vos besoins ?
- Est-il possible d'exporter des modèles de données et transformations d'une plateforme à une autre ou devrez-vous avoir recours à la rétro-ingénierie pour les recréer si vous décidez d'adopter un nouvel outil ?

Garantir l'évolutivité est essentiel, car changer de plateforme peut s'avérer très coûteux et perturbateur pour votre activité.

CONSEIL : *Si le langage SQL comprend un certain nombre de dialectes, il s'agit d'un standard dans le domaine de l'analyse. Les modèles de données et transformations écrits en langage SQL peuvent en principe être facilement transmis d'un système à un autre. En revanche, les procédures, modèles de données et transformations stockés dans des systèmes de fichiers ou écrits dans des langages propriétaires ne peuvent être transférés et impliquent un risque important d'enfermement propriétaire.*

N'OUBLIEZ PAS : *Pour évaluer la viabilité économique d'un outil, considérez son coût total de possession et sa capacité à vous épargner d'éventuelles complications organisationnelles en termes d'évolution des besoins, de résilience aux accidents et pannes et de conformité réglementaire.*

Chapitre 5 : Considérations techniques pour choisir un outil d'intégration de données

DANS CE CHAPITRE :

- Processus ETL et ELT
- Évaluer la qualité des connecteurs de données
- Comprendre comment l'automatisation fonctionne au sein de votre pile

Une fois que vous avez identifié les besoins métier auxquels doit répondre une solution d'intégration de données, il convient d'évaluer les caractéristiques techniques de chaque outil.

Qualité des connecteurs de données

Le composant de base de chaque pipeline de données ELT est le connecteur de données. Un connecteur de données ingère les données depuis une API ou un journal de base de données, applique un processus de nettoyage et de normalisation de surface, puis charge les données dans un entrepôt. Lors de l'évaluation d'un connecteur de données, prenez en compte les éléments suivants :

- **Open source et propriétaire.** À l'instar des autres composants logiciels, un compromis est nécessaire entre talent volontaire participatif et attention professionnelle dédiée. Globalement, les connecteurs open source couvrent une plus grande variété de sources de données, mais leurs équivalents propriétaires

sont souvent de meilleure qualité et s'intègrent plus facilement aux autres composants d'une pile de données. Par ailleurs, les fournisseurs de technologies propriétaires appliquent des principes stricts en matière d'assurance qualité, de maintenance et d'ingénierie.

- **Schémas standardisés et normalisation.** Les données des flux d'API ne sont généralement pas délivrées sous forme normalisée. La normalisation favorise l'intégrité des données en éliminant la redondance et en établissant des associations claires et cohérentes entre les tables. Pour un jeu de données spécifique, il existe autant d'opinions que de schémas possibles, mais seulement une poignée de schémas normalisés. Les méthodes de normalisation des jeux de données étant limitées, cette approche est également propice à la standardisation des schémas, qui offre des économies d'échelle profitant à tous les utilisateurs.

CONSEIL : *Recherchez des diagrammes entité-association (EA) illustrant les schémas dans la documentation fournisseur. Les EA doivent clairement indiquer les champs disponibles dans chaque table, ainsi que les associations entre chacune d'elles. Vos analystes doivent être en mesure de déterminer si le schéma contient des champs utiles et s'il est normalisé.*

- **Mises à jour incrémentielles et intégrales.** Quelle est la stratégie de réplication du connecteur ? La synchronisation initiale nécessitera d'interroger l'intégralité ou un sous-ensemble substantiel du jeu de données, mais pas les mises à jour consécutives. Le connecteur est-il mis à jour de façon incrémentielle à l'aide de journaux ou d'autres formes de détection des modifications ou interroge-t-il l'ensemble des jeux de données à chaque synchronisation ? Les mises à jour incrémentielles permettent des réplications plus fréquentes à faible volume. La réplication fréquente des bases de données opérationnelles dans leur ensemble présente un risque supplémentaire : elle est susceptible d'interférer avec vos opérations métier critiques.

Prise en charge des sources et destinations

Divers outils de pipeline de données prennent en charge diverses sources et entrepôts de données. Assurez-vous que l'outil que vous évaluez prend en charge ceux qui vous importent. Dans le cas contraire, le fournisseur propose-t-il aux clients un mode de suggestion de nouvelles sources et destinations ? En ajoute-t-il régulièrement de nouvelles ?

À cette fin, l'outil prend-il en charge plusieurs sources et destinations ? À terme, il se peut que votre entreprise utilise des dizaines, voire des centaines de connecteurs pour un même type de sources de données si par exemple, vous gérez de nombreux

comptes publicitaires pour vos clients ou si votre entreprise fait l'objet d'une fusion ou d'une acquisition et doit combiner les données de plusieurs comptes et plateformes. Vous pouvez également opter pour une synchronisation avec plusieurs entrepôts de données à des fins de redondance.

Enfin, déterminez si et à quel degré l'outil prend en charge les intégrations de données personnalisées. Il se peut que vous deviez intégrer les données depuis des sources obscures non prises en charge par les connecteurs standard du marché.

L'outil que vous évaluez prend-il en charge les fonctions cloud vous permettant de combiner des connecteurs personnalisés écrits par vos ingénieurs avec le reste de votre infrastructure ? En bref, l'outil prend-il en charge le chargement et l'entreposage de données ad hoc depuis des fichiers CSV ou JSON ?

Configuration et modèle « Zero-Touch »

Les outils hautement personnalisables et configurables permettent aux utilisateurs d'ajuster chaque paramètre et de définir précisément le flux de travail souhaité. Cette approche nécessite des ingénieurs qualifiés en langages de scriptage, dotés d'une profonde expérience de l'orchestration et capables de développer des logiciels robustes. Elle implique également qu'ils comprennent en détail chaque source de données ou qu'ils collaborent étroitement avec les analystes pour explorer, appréhender et modéliser les données. À terme, les schémas doivent devenir des modèles de données exploitables ; néanmoins, la mise en œuvre de schémas et d'un processus d'affinage des données viables est une tâche complexe qui relève autant de l'art que de la science.

Avec une approche hautement configurable, les utilisateurs doivent correctement paramétrer et gérer le logiciel d'intégration. Cela implique la reconfiguration des pipelines à chaque modification des besoins métier en aval et des sources de données en amont. L'approche hautement configurable est plus adaptée aux entreprises dotées d'un panel de talents hautement techniques, souhaitant traiter activement ces problématiques et réellement confiantes dans leur capacité à délivrer une solution plus fiable que les produits commerciaux.

ATTENTION : *Il existe également des outils d'intégration orientés interface utilisateur permettant au personnel sans compétences en ingénierie de programmer visuellement des orchestrations et transformations. Au lieu d'une équipe d'ingénieurs hautement qualifiés, vous aurez besoin d'analystes ou d'utilisateurs finaux dotés d'une expérience approfondie des langages de programmation visuelle propriétaires. Cette approche peut aboutir à de sérieux problèmes en termes de compétences surspécialisées ou d'enfermement propriétaire.*

En revanche, reposant sur une approche de configuration unique, les outils « Zero-Touch » entièrement gérés sont extrêmement accessibles. Du point de vue client, les connecteurs sont standardisés, éprouvés et ne nécessitent aucune maintenance. La maintenance et les itérations futures incombent aux experts comprenant toutes les spécificités des données sous-jacentes et ayant testé leurs connecteurs sur une plage étendue de cas marginaux.

Au lieu d'orchestrer et de transformer les données avant leur chargement, les transformations peuvent être planifiées et réalisées par des analystes via un système SQL. Pour cette raison, l'approche « Zero-Touch » est bien plus adaptée aux entreprises n'ayant pas accès à un panel d'ingénieurs de pointe pour développer et gérer les pipelines, et souhaitant affecter leurs talents d'ingénierie à d'autres projets à plus forte valeur.

CONSEIL : *Le libre-service est également un aspect essentiel à prendre en compte. Déterminez si et dans quelle mesure l'outil que vous évaluez permet de créer un compte sans l'intervention d'un gestionnaire de compte ou d'un membre de l'équipe de support client. Le libre-service peut légèrement augmenter la charge de travail de votre équipe, mais également vous permettre de souscrire et d'interrompre plus facilement votre abonnement.*

Automatisation

Le but des outils d'intégration de données modernes est d'éliminer au maximum les interventions et efforts manuels. À cette fin, envisagez les économies de main d'œuvre offertes par les outils et fonctionnalités d'automatisation suivants :

- **API.** Contrôler les outils par voie programmatique afin d'exécuter automatiquement les fonctions d'administration et tâches de routine au lieu de les traiter manuellement peut grandement simplifier les choses. Ces fonctionnalités sont particulièrement utiles lorsqu'un grand nombre de personnes nécessitent différents niveaux de contrôle sur l'outil ou si vous développez des produits basés sur l'intégration de données.
- **Gestion des modifications de type de données.** La modification des schémas en amont peut altérer le type d'une valeur particulière ; par exemple, transformer un nombre entier en valeur flottante. Un outil automatisé doit être en mesure de réconcilier les anciens et nouveaux types de données sans intervention humaine.
- **Planification de la synchronisation continue.** Les données issues de ces connecteurs doivent être transmises à votre entrepôt ou synchronisées à intervalles très réguliers. Déterminez la fréquence de mise à jour des données dont votre entreprise a besoin, plus configurez-la une fois pour toutes.

- **Migrations de schémas automatiques.** Les schémas changeront inévitablement lors de l'ajout de nouveaux éléments au jeu de données. Le connecteur s'adapte-t-il automatiquement à ces changements avec un impact minimal sur les éléments en aval (c.-à-d., sans supprimer de tables ou de champs) ? Le connecteur évite-t-il les resynchronisations intégrales dès que possible ?
- **Performances globales.** Il convient enfin de prendre en compte un certain nombre de caractéristiques qui détermineront les temps d'arrêts éventuels de votre système, notamment les suivantes :

    - Combien de temps prend la synchronisation initiale ?
    - Les données sont-elles mises à jour de façon incrémentielle, ou une synchronisation intégrale est-elle nécessaire à chaque fois ?
    - Quelles conditions déclenchent une synchronisation intégrale ?
    - Quelle est la fréquence de mise à jour des données et dans quelle mesure répondelle à vos besoins ?

        - Les flux de données sont-ils diffusés en direct ? À quelques minutes d'intervalle ? Une fois par jour ?

Les réponses à ces questions peuvent influencer les coûts imposés par les temps d'arrêt et l'exploitation de l'infrastructure — ces coûts ne seront sans doute pas inclus dans la structure de tarification formelle de l'outil, mais peuvent avoir un impact significatif pour votre entreprise.

Transformations d'entrepôt intégrées et processus antérieurs

Avec le processus ELT, les transformations sont exécutées dans un entrepôt de données cloud élastique. L'élasticité — et la séparation entre capacités de traitement et de stockage — permet d'adapter les ressources selon les besoins. Cette approche élimine la nécessité de prévoir les exigences matérielles et d'acheter une capacité excédentaire.

En revanche, le processus ETL, qu'il ait lieu dans le cloud ou sur site, nécessite une architecture de données comprenant une étape supplémentaire dans la pile de données, afin de gérer les transformations avant leur chargement. Si la pile de données se trouve sur site, l'entrepôt de données peut contraindre le volume de données chargées, ce qui implique des transformations visant à limiter le volume et le flux de données.

L'avantage fondamental du processus ELT et des transformations appliquées dans l'entrepôt de données est qu'ils sont non destructifs — c'est-à-dire qu'ils conservent les données sous-jacentes intactes lors de la création de tables supplémentaires intégrant les modèles souhaités. Cela signifie que si elles échouent, les tentatives de transformation n'ont aucune conséquence permanente et qu'elles sont réitérables.

Par ailleurs, les analystes peuvent ajuster les modèles à l'évolution des besoins métier sans perte de données.

Dernier avantage de l'application des transformations dans l'entrepôt de données : les transformations peuvent être écrites en langage SQL, ce qui les rend accessibles aux analystes. Ces derniers créent souvent des vues dans les entrepôts afin de consolider ou de modifier les tables. Les outils ELT prenant en charge les transformations d'entrepôt intégrées permettent aux analystes de créer des vues de manière systématique.

Récupération après panne

Des bogues et erreurs apparaîtront inévitablement au cours du processus d'intégration de données et certaines instances échoueront inéluctablement. Dans ce cas, vous souhaiterez à tout prix éviter de perdre des données de façon permanente.

L'une des caractéristiques essentielles d'un outil d'intégration de données est l'*idempotence* — la capacité à réitérer un même processus en produisant le même résultat à chaque fois. Elle est particulièrement importante dans le cas de processus complexes en plusieurs étapes, pour lesquels le point de rupture n'est pas nécessairement évident.

L'*intégration additive nette* est un autre principe important. Lorsqu'une valeur est supprimée dans les données source ou qu'une table est éliminée, est-elle conservée (mais marquée) dans l'entrepôt de données ? La conservation avec marquage des valeurs supprimées préserve les registres historiques et favorise les opérations d'audit, de récupération après panne, ainsi que l'analyse des tendances et de l'attrition à long terme.

CONSEIL : *N'oubliez pas de lire les accords de niveau de service (SLA) de chaque fournisseur que vous évaluez et de veiller à leur application. Plus particulièrement, assurez-vous que les SLA du fournisseur sont à la hauteur de ceux que vous exigez de votre équipe. Un SLA définit des attentes claires concernant la disponibilité, les temps d'arrêt, le débit et le volume de transfert des données, ainsi que d'autres mesures de performances.*

Conformité aux exigences réglementaires et de sécurité

La cybersécurité et la confidentialité sont des problématiques hautement sensibles, d'un point de vue légal et public.

Voici quelques considérations à ne pas négliger :

- **Conformité réglementaire.** Votre fournisseur d'intégration de données doit au minimum être au fait des normes telles que RGPD, SOC 2, HIPPA et autres réglementations applicables. Un outil pertinent doit prendre en charge l'omission ou le chiffrement des informations personnellement identifiables (IPI).
- **Propriété de vos données.** Votre fournisseur d'intégration de données ne doit pas avoir accès à vos données ni les conserver plus longtemps que nécessaire pour les répliquer.
- **Rôles à niveaux d'accès variables.** Tous les utilisateurs de l'outil ne doivent pas disposer de droits illimités en ce qui concerne la création, la suppression ou la modification des entrepôts de données, connecteurs et transformations, ou encore l'exécution d'autres actions sensibles. L'outil doit fournir un éventail de rôles de l'administration à la lecture seule.
- **Blocage et hachage de colonnes.** Afin de garantir la sécurité et la conformité réglementaire, vous devez être en mesure d'obscurcir ou d'omettre les IPI dans chaque table que vous synchronisez.

N'OUBLIEZ PAS : *Le choix d'un outil d'intégration de données dépend fondamentalement de sa capacité à faciliter le travail des analystes et ingénieurs.*

Considérez les points suivants :

- *Qualité des connecteurs de données*
- *L'outil prend-il en charge les sources et destinations de données que vous utilisez ou prévoyez d'utiliser ?*
- *Quel volume de configuration pratique nécessite-t-il ?*
- *Quel niveau de fonctionnement offre-t-il sans intervention manuelle ou supervision ?*
- *À quel moment l'outil exécute-t-il les transformations ?*
- *Quelle est la résilience de l'outil aux pannes ?*
- *Conformité aux exigences réglementaires et de sécurité*

Essayez plusieurs outils ! La section suivante décrit les différentes étapes de démarrage.

Chapitre 6 : Démarrage en sept étapes

DANS CE CHAPITRE :

- Comprendre vos besoins et objectifs
- S'assurer que votre perspective de réussite est viable
- Essayer et réessayer avant d'acheter

Si les promesses d'une pile de données cloud entièrement gérée sont séduisantes, cette solution n'est pas adaptée à toutes les entreprises.

Pour choisir l'outil le plus pertinent pour votre entreprise, vous devez :

1. Évaluer rigoureusement vos besoins
2. Décider d'opter pour une migration ou une nouvelle instance
3. Évaluer les entrepôts de données cloud et outils d'informatique décisionnelle
4. Évaluer les outils d'intégration de données
5. Calculer le coût total de possession
6. Établir les critères de réussite
7. Définir une preuve de concept

Évaluation des besoins

Pour un certain nombre de raisons, il peut être préférable de ne pas externaliser vos opérations de données à un tiers ou dans le cloud.

La première et la plus évidente est que vous opérez dans une PME ou avec un niveau de complexité très faible en matière de données. Si par exemple, vous travaillez dans une start-up de quatre personnes en quête du segment de marché pertinent, il se peut même que vous n'ayez aucun processus impliquant des données. C'est également le cas si vous n'utilisez qu'une ou deux applications, ne prévoyez pas d'en adopter de nouvelles et si les outils d'analyse intégrés à vos applications vous suffisent.

La deuxième raison pour laquelle il est sans doute préférable de ne pas adopter une pile de données moderne est qu'elle peut être incompatible avec certaines normes de performance ou de conformité réglementaire. Si vous opérez au sein d'une société commerciale intensive et que quelques nanosecondes suffisent à impacter vos opérations, vous souhaiterez sans doute éviter les infrastructures cloud tierces au profit d'une solution développée en interne.

En revanche, si l'envergure ou la maturité de votre entreprise est suffisante pour tirer parti des technologies d'analyse de données, et que des cycles d'actualisation de quelques minutes, voire de quelques heures sont acceptables, alors n'hésitez pas.

Migration ou nouvelle instance

Les fournisseurs de solutions d'intégration des données doivent être en mesure de migrer vos données depuis l'ancienne infrastructure vers votre nouvelle pile de données ; notez néanmoins qu'il s'agit d'une initiative très fastidieuse en raison de la complexité intrinsèque et de la diversité des données. La décision de votre entreprise d'opter pour une migration ou une nouvelle instance repose principalement sur la valeur estimée des données historiques.

Si votre entreprise a déjà acheté ou souscrit des produits ou services, résilier ces contrats peut s'avérer onéreux. Outre l'aspect financier, il peut être important de prendre en compte la familiarisation et les préférences de votre équipe pour certains outils.

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

CONSEIL : *Une approche échelonnée est tout à fait normale. De nombreuses entreprises mettent en œuvre un nouvel entrepôt de données tout en conservant l'ancien à des fins de redondance, jusqu'à ce que l'ensemble des données et processus aient été migrés vers le nouvel environnement.*

Évaluation des entrepôts de données cloud et outils d'informatique décisionnelle

Vous devrez évaluer et comparer des solutions pour chaque composant de la pile de données. Avant d'acquérir un outil d'intégration de données, descendez un peu plus en aval et réfléchissez aux fonctionnalités que vous attendez d'un entrepôt de données et d'un outil de BI.

Voici quelques-unes des fonctionnalités des entrepôts de données cloud à prendre en compte :

1. Stockage de données centralisé ou décentralisé
2. Élasticité – l'entrepôt de données offre-t-il une mise à l'échelle rapide des ressources ? Les ressources de traitement et de stockage sont-elles indépendantes ou étroitement liées ?
3. Charges concurrentes– l'entrepôt de données peut-il gérer plusieurs tâches simultanément ?
4. Performances des processus de chargement et d'interrogation
5. Gouvernance des données et gestion des métadonnées
6. Dialecte SQL
7. Prise en charge du processus de sauvegarde et de restauration
8. Résilience et disponibilité
9. Sécurité

Voici quelques-unes des fonctionnalités des outils de BI à prendre en compte :

1. Simplicité d'intégration aux entrepôts de données cloud
2. Interfaces de type glisser-déposer simples d'utilisation – particulièrement utiles si vous souhaitez développer une culture orientée données à l'échelle de l'entreprise
3. Reporting et notifications automatisés
4. Capacité à réaliser des calculs et à créer des rapports ad hoc en ingérant et en exportant les fichiers de données
5. Vitesse, performance et réactivité
6. Couche de modélisation avec contrôle de version et mode de développement

7. Bibliothèque de visualisations exhaustive

Assurez-vous que les entrepôts de données et outils de BI que vous évaluez sont intercompatibles. Il peut également être pertinent d'examiner diverses perspectives sur différents outils.

Les publications telles que le rapport Gartner rassemblent souvent des informations de ce type. Veillez à bien vous documenter avant de vous décider.

Évaluation des outils d'intégration de données

Comme nous l'avons vu plus haut, de nombreuses caractéristiques sont à prendre en compte concernant les outils d'intégration de données.

En bref, les principaux aspects à évaluer sont les suivants :

1. Personnalisation et configurabilité / simplicité d'utilisation et accessibilité
2. Fiabilité et performances du logiciel
3. Qualité et réactivité des équipes de support client
4. Nombre et type de sources de données pris en charge
5. Coûts et plans de paiement

De nombreuses publications offrent des études et évaluations agrégées portant sur les outils d'intégration de données comme sur les entrepôts de données et outils de BI. Comparez attentivement les produits avant d'acheter.

Assurez-vous que les outils d'intégration de données que vous étudiez sont compatibles avec les entrepôts de données et outils de BI dont vous disposez ou que vous prévoyez d'acquérir.

Calcul du coût total de possession et du retour sur investissement

La pile de données moderne offre des économies substantielles en termes de temps, d'argent et de main-d'œuvre. Comparez votre flux de travail d'intégration de données actuel à un éventail de produits prospects.

Calculez le coût de votre pipeline de données actuel, ce qui nécessitera sans doute un audit strict des dépenses antérieures en activités d'intégration de données.

Vous devrez prendre en compte le prix d'achat, les coûts de configuration et de maintenance, ainsi que les coûts ponctuels liés aux pannes, interruptions et temps d'arrêt. Enfin, n'oubliez pas d'inclure le coût de votre entrepôt de données et de votre outil de BI.

En second lieu, vous devrez évaluer les avantages de la solution de remplacement envisagée. Si certains ne seront sans doute pas très tangibles ou calculables (p. ex. amélioration potentielle de la motivation des analystes), d'autres, tels que les gains de temps et d'argent, sont aisément quantifiables.

Établissement des critères de réussite

À quoi le flux de travail de vos analystes doit-il ressembler si vous avez correctement mis en œuvre une pile de données moderne ?

Les critères clés comprennent notamment les suivants :

1. Gains de temps, d'argent et de main-d'œuvre par rapport à la solution précédente
2. Développement des capacités de l'équipe de données
3. Mise en œuvre réussie des nouveaux projets de données, tels que les modèles d'attribution client
4. Réduction des délais de traitement et de création de rapports
5. Réduction des temps d'arrêt de l'infrastructure de données
6. Adoption croissante de l'outil de BI au sein de votre entreprise
7. Disponibilité et exploitabilité de nouvelles métriques

Définition d'une preuve de concept

Après avoir restreint votre recherche à quelques outils prospects et déterminé les standards de réussite, testez les produits sur des projets dotés d'un enjeu négligeable. La plupart des produits offrent des essais gratuits de quelques semaines.

Configurez les connecteurs entre vos sources de données et votre entrepôt, puis mesurez le temps et les efforts requis pour synchroniser vos données et appliquez des transformations de base. Aménagez une plage d'essai pour votre équipe et encouragez celle-ci à tester les performances du système de toutes les façons possibles.

Comparez les résultats de votre essai à vos standards de réussite.

N'OUBLIEZ PAS : *L'intégration de données automatisée peut considérablement améliorer les capacités de vos analystes et ingénieurs des données, mais il est impératif de comprendre toute l'étendue de vos besoins et objectifs avant de commencer. Développez une perspective de réussite (ou d'échec) et assurez-vous de tester rigoureusement les performances de la pile de données évaluée avant de l'adopter.*

L'analyse compétitive commence par l'intégration de données automatisée

L'accroissement des données cloud a généré des opportunités sans précédent en termes de développement produit et d'analyses avancées. Le volume et la complexité de ces données représentent néanmoins un défi en matière d'intégration des données que peu d'entreprises sont en mesure de relever. L'intégration de données automatisée aide les entreprises à exploiter pleinement les données pour accélérer et améliorer la prise de décisions stratégiques. Ce guide retrace l'historique de l'intégration de données, évalue les solutions actuelles et vous explique comment choisir l'outil d'intégration de données le mieux adapté à votre entreprise.

Dans cet ouvrage :

- L'évolution de la pile de données moderne
- Pourquoi le processus d'intégration de données est-il encore plus complexe qu'auparavant
- Pourquoi l'automatisation est-elle essentielle pour l'intégration de données moderne
- Créer ou acheter une solution d'intégration de données
- Critères métier et techniques pour choisir une solution d'intégration de données
- Comment mettre en œuvre une solution d'intégration de données

Charles Wang, Évangéliste produit chez Fivetran, a précédemment opéré en tant qu'analyste des données, expert des données et responsable produit.

ISBN 978-1-7346299-0-3

90000>

9 781734 629903

À propos de Fivetran

Leader du secteur de l'intégration de données automatisée, Fivetran délivre des connecteurs, transformations et modèles d'analyse prêts à l'emploi. Les connecteurs Fivetran acheminent les données en continu vers un référentiel central et s'adaptent aux modifications de schémas et d'API. Cette approche garantit un accès fluide et fiable aux données, ce qui vous permet d'adopter toutes les applications SaaS dont vous avez besoin et de poursuivre vos efforts d'analyse en toute confiance. Pour en savoir plus sur l'avenir de l'intégration de données, rendez-vous sur fivetran.com.