



talend

Guide Complet De La Gouvernance Des Données

Sommaire

Introduction : La fiabilité des données conditionne la transformation numérique.	03
Chapitre 1 : Qu'est-ce que la gouvernance des données ? Et pourquoi en avez-vous besoin ?	06
Chapitre 2 : Choisir le modèle de gouvernance le plus adapté	13
Chapitre 3 : Trois étapes pour fournir rapidement des données fiables.	21
Chapitre 4 : Recommandations : les 12 travaux de la gouvernance des données	44
Chapitre 5 : Les nouveaux rôles de la gouvernance des données.	50
Chapitre 6 : Exemple de réussite en gouvernance de données	54
Chapitre 7 : De l'intégration à l'intégrité des données	65
Chapitre 8 : Devenir une entreprise « data intelligente »	72

Introduction



La fiabilité des données conditionne la transformation numérique

À l'ère de l'économie de l'information, les données constituent les ressources les plus précieuses des entreprises. Dans tous les secteurs, les stratégies orientées données constituent un impératif concurrentiel. Pour atteindre leurs objectifs de croissance, de rentabilité et de satisfaction client par exemple, les entreprises s'appuient de plus en plus sur les données pour prendre leurs décisions. Ce processus décisionnel piloté par les données est essentiel à toute initiative de transformation numérique.

Mais, pour disposer des données nécessaires à leur transformation numérique, les entreprises doivent concilier deux paramètres essentiels.

Caractérisée par la rapidité et l'accélération de la commercialisation pour répondre en temps réel aux demandes des équipes commerciales ou offrir des expériences personnalisées aux clients, la transformation numérique exige des données disponibles en temps voulu. Toutefois beaucoup d'entreprises ne parviennent pas encore à suivre le rythme soutenu des projets technologiques.

Néanmoins, si la rapidité est essentielle, elle ne fait pas tout. Les entreprises doivent être assurées de la fiabilité de leurs données pour permettre une prise de décisions efficace et pour offrir des expériences client exceptionnelles. Ceci représente un défi majeur pour les entreprises. Et c'est bien cette garantie qui permet aux entreprises de conserver la confiance de leurs clients dans le respect de la réglementation et d'assurer que les bonnes personnes ont accès aux données pertinentes pour prendre des décisions judicieuses. Un défi majeur pour les entreprises. Selon la Harvard Business Review, [47 % des enregistrements de données](#), en moyenne, contiennent des erreurs critiques, qui ont des répercussions sur l'activité de l'entreprise.

D'après Forrester,
seuls **40 %** des
directeurs informatiques
fournissent des résultats
dans les délais attendus.

Rapidité et fiabilité ne faisant généralement pas bon ménage, il n'est pas rare que les entreprises privilégient l'une ou l'autre. Nombre d'entre elles misent par commodité sur la rapidité pour répondre aux attentes des utilisateurs des données et respecter leurs impératifs. Pour obtenir des résultats rapides, les développeurs sont autorisés à coder manuellement les intégrations ou à travailler sur des projets ponctuels avec des outils d'intégration très spécialisés. Ces astuces fonctionnent sur le court terme, mais ne peuvent pas évoluer avec l'entreprise. En outre, le manque de supervision engendre des risques en matière de qualité et conformité. À l'inverse, les entreprises qui s'efforcent de remédier au problème de la fiabilité des données mettent souvent en place des contrôles très stricts et une gouvernance rigoureuse au point d'entraver toute possibilité d'action. Mais cette stratégie fastidieuse, contraignante et lente accapare également de nombreuses ressources. Ceci représente un véritable frein à l'innovation et l'agilité dont les entreprises ont tant besoin pour se démarquer. Le risque pour les entreprises qui manquent de réactivité est de se faire rapidement distancer.

Il est possible de concilier rapidité et fiabilité sans avoir à faire de compromis. Pour offrir en permanence une fiabilité des données, les entreprises doivent les gérer de façon globale, tout au long de leur cycle de vie, et assurer le respect de la réglementation et la sécurité des données, sans trop restreindre l'accès aux données et leur utilisation.

Grâce à notre nouvelle plate-forme d'analyses des données, nous pouvons désormais mieux comprendre l'évolution du marché, ce qui nous aide à optimiser le trading énergétique tout en gérant les risques et en respectant les réglementations. »

René Greiner, Vice President for Data Integration chez Uniper SE

Chapitre 1 :

Qu'est-ce que la gouvernance
des données ? Et pourquoi
en avez-vous besoin ?

Pourquoi moderniser votre approche des données

Imaginez que vous recherchez désespérément un livre rare. La seule possibilité pour l'obtenir est de vous rendre dans une bibliothèque et il n'y en a qu'une seule dans votre ville. Un contrôle strict est exercé à l'entrée et vous devez présenter votre carte d'identité pour obtenir une carte d'accès. Une fois à l'intérieur, vous devez vous frayer un chemin à travers des rayonnages de livres très serrés. Vous vous rendez vite compte qu'il sera difficile de retrouver votre livre, car rien n'est rangé dans cette pagaille. Aucun des rayonnages n'est classé par titre ou auteur. Vous continuez malgré tout à chercher. Comme rien n'est indiqué, vous devez regarder chaque livre pour voir si c'est le bon. Vous pourriez demander de l'aide à un bibliothécaire, mais il est probablement trop occupé à réceptionner de nouveaux livres ou à satisfaire d'autres visiteurs pour vous aider.

Au bout d'un certain temps, vous trouvez enfin le précieux livre. Mais quand vous l'ouvrez, vous découvrez que certaines pages ont été déchirées : le livre est donc difficile à comprendre et ne présente plus aucun intérêt pour vous.

Ne blâmez pas le bibliothécaire ; il doit aussi s'occuper des CD et des DVD, du classement des nouveaux formats numériques et d'une file d'attente croissante (ainsi que des visiteurs en ligne réclamant des références supplémentaires).

Vous aurez peut-être quelques idées pour améliorer l'organisation afin qu'il soit possible de trouver plus rapidement les livres. Mais personne n'a demandé votre aide ; vous n'étiez là qu'en tant que lecteur. En outre, l'intégrité globale de cette bibliothèque n'inspire pas vraiment confiance. Les conditions déplorables, la mauvaise qualité des livres et le temps précieux que vous avez perdu vous en donnent une mauvaise image ; il ne s'agit de toute évidence pas d'un établissement sérieux que vous recommanderiez.

Cette expérience vous semble décourageante et frustrante ? La communauté d'utilisateurs de vos données peut ressentir la même chose quand elle est amenée à rechercher des ensembles de données dans votre entreprise.

Dans une entreprise comme dans une bibliothèque, nous devons gérer un volume croissant de données, provenant des sources de données traditionnelles utilisées par le passé ou générées par l'ère du numérique (par exemple réseaux sociaux, données de capteurs, etc.).

Ceci entraîne une prolifération des données dont il est presque impossible de mesurer l'ampleur. Plus vous recueillez des données, moins vous pouvez garantir le libre-service. Votre bibliothèque de données n'est utile qu'aux quelques privilégiés disposant du large socle de compétences nécessaire pour découvrir par eux-mêmes la valeur qu'elles recèlent. Les autres ne sauront pas en tirer parti.

Les énormes volumes de données arrivant de partout se traduisent également par une perte de contrôle. Il se peut que vous ne soyez même pas conscient que les données que vous recevez et mettez à la disposition d'autrui contiennent des informations inappropriées ou inexactes et par conséquent peu fiables. Nous observons cette « prolifération des données » dans de nombreuses entreprises qui se retrouvent incapables de faire face au rythme et au volume de données entrant dans leurs systèmes.

Imaginez que nous puissions rendre toutes ces données fiables, les organiser à grande échelle et les fournir à tous ceux qui en ont besoin ? C'est-à-dire offrir à vos collaborateurs les outils leur permettant de fiabiliser, d'organiser et de distribuer les données par eux-mêmes. Cette capacité est l'essence même de la gouvernance des données.

« En réunissant nos données dans un seul système, nous améliorons la qualité et facilitons la gouvernance des données. »

Responsable informatique, fournisseur de services de télécommunication aux entreprises

Qu'est-ce que la gouvernance des données ?

Outre le contrôle et la protection des données, la gouvernance des données, englobe également les informations obtenues par le biais de l'intégration et du mode participatif (crowdsourcing). La gouvernance des données est une nécessité dans l'environnement d'entreprise actuel hautement compétitif et en rapide mutation. Les entreprises ont aujourd'hui la possibilité de collecter d'énormes quantités de données internes et externes très diverses. Elles doivent par conséquent pouvoir s'organiser pour optimiser la valeur de ces données, gérer les risques associés et réduire les coûts de gestion.

La gouvernance des données regroupe des processus, des rôles, des politiques, des normes et des mesures qui garantissent une utilisation efficace de l'information pour permettre à une entreprise d'atteindre ses objectifs. Elle définit les processus et responsabilités qui garantissent la qualité et la sécurité des données utilisées au sein d'une entreprise.

Une stratégie de gouvernance des données bien conçue est vitale pour toute entreprise.

Une stratégie de gouvernance des données bien conçue est vitale pour toute entreprise manipulant des données. Elle déterminera la façon dont votre entreprise tirera profit de processus et de responsabilités normalisés uniformes. Les leviers métiers mettront en évidence les données devant être scrupuleusement surveillées dans votre stratégie de gouvernance des données et les avantages qui devraient en découler. Cette stratégie constituera le fondement de votre cadre de gouvernance des données.

Par exemple, si un levier de cette stratégie vise à garantir la confidentialité de données médicales, les données des patients devront être gérées de façon sécurisée lorsqu'elles circulent dans votre entreprise. Les exigences en matière de conservation (par exemple, historique des changements apportés aux informations, avec leur date et leur auteur) seront définies de sorte à garantir la conformité avec les exigences gouvernementales applicables ([RGPD](#) en Europe ou [CCPA](#) en Californie).

La gouvernance des données garantit que les rôles sont clairement définis et que les responsabilités font l'objet d'un consensus dans toute l'entreprise. Un cadre de gouvernance des données planifié avec soin englobe les rôles et les responsabilités stratégiques, tactiques et opérationnels.

La gouvernance des données est incontournable

Une stratégie de gouvernance des données efficace présente tellement d'avantages essentiels pour votre entreprise qu'elle aura du mal à s'en passer.

Principaux avantages :

- **Une compréhension commune des données :** la gouvernance des données offre une vision uniforme et une terminologie commune pour les données, tout en laissant une latitude adéquate aux différents départements de l'entreprise.
- **Amélioration de la qualité des données :** la gouvernance des données met en place un plan garantissant l'exactitude, l'exhaustivité et la cohérence des données.
- **Une cartographie des données :** la gouvernance des données permet de connaître précisément l'emplacement de toutes les données liées à des entités critiques, une étape nécessaire à l'intégration des données. Tel un GPS qui restitue la géographie des lieux et aide ses utilisateurs à retrouver leur chemin en territoire inconnu, la gouvernance des données rend les données exploitable et permet d'établir plus facilement un lien avec les résultats opérationnels.
- **Une vue à 360° de chaque client et des autres entités commerciales :** la gouvernance des données établit un cadre permettant à l'entreprise de définir un référentiel unique pour les entités commerciales critiques. Elle peut ensuite instaurer un niveau d'uniformité approprié dans l'ensemble de ses entités et activités.
- **Une uniformité de la conformité :** la gouvernance des données prépare le terrain de la conformité aux réglementations gouvernementales, par exemple le Règlement général sur la protection des données (RGPD) en Europe, la loi de protection des consommateurs (California Consumer Protection Act, CCPA) en Californie, la loi américaine sur l'assurance maladie (Health Insurance Portability and Accountability Act, HIPAA), et à des exigences sectorielles comme les normes de sécurité des données applicables à l'industrie des cartes de paiement (Payment Card Industry Data Security Standards, PCI DSS).
- **Une gestion améliorée des données :** la gouvernance des données apporte une dimension humaine dans un monde automatisé où les données sont reines. Elle instaure des codes de conduite et des bonnes pratiques en matière de gestion des données afin de répondre de manière cohérente aux préoccupations et aux besoins dans les domaines les plus novateurs (droit, sécurité et conformité, notamment).
- **Une facilité d'accès :** en instaurant un cadre de gouvernance des données dans votre entreprise, vos données seront fiables, bien documentées et faciles à trouver et vous garantirez leur sécurité, leur conformité et leur confidentialité.

Une stratégie de gouvernance des données efficace présente tellement d'avantages essentiels pour votre entreprise qu'elle aura du mal à s'en passer.

RGPD, CCPA... : La gouvernance des données pour répondre aux enjeux réglementaires

L'Union européenne (UE) a publié le [Règlement général sur la protection des données](#) (RGPD) en mai 2016. Après une période de transition de deux ans, le RGPD est entré en vigueur le 25 mai 2018. Le RGPD concerne le traitement des données personnelles de toutes les personnes concernées, y compris les clients, les salariés et les prospects. La définition des données personnelles comprend « toute information se rapportant à une personne physique identifiée ou identifiable (ci-après dénommée "personne concernée") ; est réputée être une "personne physique identifiable" une personne physique qui peut être identifiée, directement ou indirectement, notamment par référence à un identifiant, tel qu'un nom, un numéro d'identification, des données de localisation, un identifiant en ligne, ou à un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale ». Parmi les exemples types, on peut citer les noms et les coordonnées de clients dans le cadre d'un CRM, ainsi que les salaires, avantages et performances des salariés, mais ceci s'applique aussi aux types de données plus récents tels que les données de capteurs permettant d'indiquer la localisation des véhicules ou le comportement des conducteurs.

La réglementation s'applique aux personnes concernées dans l'Union européenne, même lorsque les données sont traitées par des entreprises opérant en dehors de l'UE dans des juridictions comme les États-Unis, l'Asie Pacifique, le Moyen-Orient et l'Afrique. Le non-respect du RGPD peut donner lieu à des amendes très lourdes pouvant s'élever à 20 millions d'euros ou 4 % du chiffre d'affaires mondial de l'entreprise, le montant le plus élevé étant celui retenu. Le RGPD n'est pas la seule grande législation en matière de protection des données. L'État de Californie lui a emboîté le pas avec la [CCPA](#) (California Consumer Privacy Act) qui confère aux consommateurs le droit de savoir quelles informations les entreprises recueillent les concernant, pourquoi elles recueillent ces données et avec qui elles les partagent.

Un solide programme de gouvernance des données est une pièce maîtresse de la conformité avec les législations sur la protection des données. En outre, la conformité au RGPD doit intégrer des contrôles en libre-service liés à la préparation et à la gestion des données, le but étant de favoriser la responsabilisation des acteurs en ce qui concerne la protection des données d'une façon concrètement vérifiable et pas simplement définie par des directives légales écrites. Si la conformité réglementaire est souvent à l'origine des projets de conformité en matière de protection des données, elle ne doit pas en être le seul moteur. L'objectif est plutôt d'instaurer des relations de confiance avec vos clients en ce qui concerne leurs données personnelles.

Une plate-forme de données de pointe pour garantir le succès d'une gouvernance des données moderne

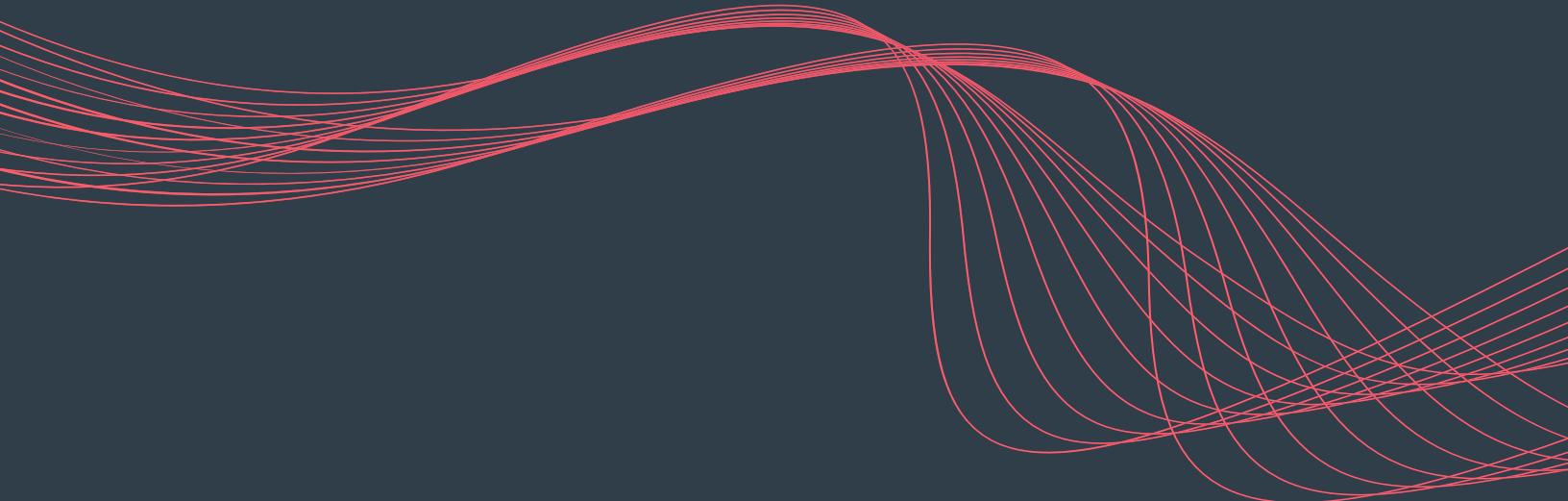
Pour trouver l'approche qui convient à votre entreprise en matière de gouvernance des données, optez pour des outils open source évolutifs faciles à intégrer dans son environnement. En outre, une plate-forme cloud vous permettra d'exploiter rapidement des fonctionnalités puissantes à la fois rentables et simples d'utilisation. Et, grâce aux solutions cloud, finis les frais indirects associés aux serveurs sur site. Au moment de comparer et de sélectionner vos outils de gouvernance des données, privilégiez ceux qui vous permettent de tirer parti des avantages définis dans votre stratégie correspondante.

Ces outils doivent vous aider à :

- Capturer et comprendre vos données grâce à des outils et des fonctionnalités de découverte, de profilage et de benchmarking. Par exemple, l'utilisation des bons outils peut permettre de détecter automatiquement des données personnelles, comme un numéro de sécurité sociale, dans un nouvel ensemble de données et de déclencher une alerte.
- Améliorer la qualité de vos données grâce à leur validation, leur nettoyage et leur enrichissement.
- Gérer vos données avec des processus ETL et ELT orientés métadonnées et des applications d'intégration de données afin de garantir la traçabilité des pipelines de données grâce au lignage des données de bout en bout.
- Contrôler vos données à l'aide d'outils assurant un examen et une surveillance actifs. Documenter vos données afin qu'elles puissent être enrichies de métadonnées et ainsi améliorer leur pertinence, leur facilité de recherche, leur accessibilité, leur chaînabilité et leur conformité.
- Permettre à ceux qui connaissent le mieux les données de contribuer à la gestion à l'aide d'outils en libre-service.

Chapitre 2 :

Choisir le modèle de gouvernance le plus adapté



Trouver le juste équilibre entre gouvernance ascendante et descendante

Le pilotage par les données est un impératif économique, mais des difficultés majeures doivent être surmontées pour tirer le meilleur parti de ces données. Les entreprises sont confrontées à une croissance exponentielle du volume de données qu'elles doivent gérer, de l'ordre de 200 pour cent tous les deux ans, ce qui se traduit par une prolifération des données. Dans le même temps, l'éventail des données à traiter et analyser (nouvelles données de streaming provenant de l'« Internet des objets », capteurs, blogs, flux de clics, données participatives des applications numériques et des réseaux sociaux, etc.) est plus vaste.

En outre, les nouveaux rôles orientés données au sein de votre entreprise se sont multipliés. Au début des années 2000, il y avait des développeurs informatiques, des analystes et des utilisateurs métiers. Puis sont apparus de nouveaux professionnels des données qui travaillaient parfois dans une organisation centrale telle qu'un département IT ou un centre d'excellence analytique et rendaient parfois compte aux différents départements, comme les data stewards, les data scientists, les data curators, les responsables de la protection des données ou les ingénieurs de données. Aujourd'hui, même les non-techniciens ont appris à maîtriser les données, aspirent à être plus que de simples consommateurs de données passifs et souhaitent transformer les données en informations de façon autonome.

Le dilemme est le suivant : des données arrivent de toute part — de sources traditionnelles déjà sous le contrôle du département IT central et de nouvelles sources de données de tous horizons (shadow IT, données tierces, « Internet des objets », applications, etc). De plus, ces données doivent être accessibles plus rapidement que jamais, maintenant que les entreprises doivent ingérer et analyser des données en temps réel au lieu de réagir à des données vieilles d'un jour ou d'une semaine. En effet, on observe

désormais un [taux de croissance annuel de l'analyse en streaming \(streaming analytics\) de 35 %](#). Et, alors qu'une entreprise est composée de nombreux salariés travaillant dans un grand nombre d'entités et avec des compétences en analyse de données très variées, le département IT est censé leur fournir un accès à l'ensemble des données à des fins de business intelligence. Son budget et ses ressources sont toutefois relativement limités. Le décalage entre les attentes de l'entreprise et ce que le département IT peut fournir ne cesse donc de croître.

De ce fait, le modèle économique de l'intégration des données est mal en point. Les organisations centrales ont traditionnellement mis en place ce qu'IDC appelle la « gouvernance par le non », ce qui signifie que les utilisateurs métiers devaient leur présenter des requêtes qu'elles étaient habilitées à satisfaire ou à rejeter. Cette situation a creusé un fossé entre l'entreprise et le département IT en matière de propriété des données, écart qui ne fait que s'accentuer face aux réalités de la prolifération des données.

Le modèle économique de l'intégration des données est mal en point.

Il y a un décalage entre les attentes de l'entreprise et ce que le département IT peut fournir.

L'approche traditionnelle de la gouvernance des données ne peut pas être adaptée à l'ère du numérique : trop peu de gens accèdent à trop peu de données.

Par le passé, nous instaurions des approches fortement centralisées et trop rigides pour gérer les données, par exemple Master Data Management, un CRM, permettant la création d'entrepôts de données pour les clients ou pour les entreprises.

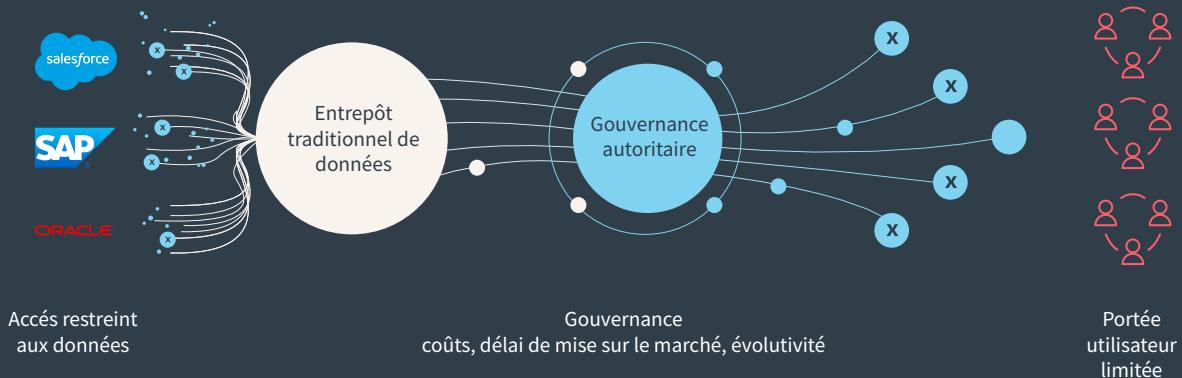
Une approche fortement centralisée s'appuie sur une petite équipe de professionnels des données hautement expérimentés dotés de méthodologies bien définies et de bonnes pratiques éprouvées. Une fois appliquée à un entrepôt de données d'entreprise, par exemple, vous pourriez commencer par définir un modèle de données central dans lequel vous pourriez recueillir et rapprocher des données identifiées comme sources d'informations. Les données seraient ensuite remaniées au sein de data marts (des bases de données connexes) de façon à pouvoir être adaptées à un domaine d'activité ou un problème. Pour finir, elles seraient remaniées une nouvelle fois à l'aide d'un outil d'informatique décisionnelle (ou business intelligence) afin d'obtenir une couche sémantique, par exemple un « catalogue de données » destiné à être intégré dans des rapports prédéfinis. Ce n'est qu'alors que les données pourraient être utilisées à des fins d'analyses.

Pour comprendre le problème d'évolutivité auquel est confronté ce modèle à l'ère du numérique, comparons-le à un domaine déjà confronté à une prolifération des données similaire : les contenus Web. Avant le 21e siècle, nous recherchions des informations dans une encyclopédie comme Encyclopædia Britannica ou Microsoft Encarta. Le modèle a créé une distinction nette entre les fournisseurs et les consommateurs de données ; seule une poignée d'experts pouvait rédiger l'encyclopédie, les autres étant des consommateurs de données.

Avec ce modèle, la qualité était excellente. L'Encyclopædia Britannica est rédigée par environ 100 éditeurs à plein temps et près de 4 000 contributeurs hautement qualifiés, dont des lauréats du prix Nobel et d'anciens chefs d'État.

Mais ce modèle a été confronté à un problème avec l'avènement du Web : ces encyclopédies n'ont pas su faire face à la demande émanant des consommateurs de données. Aujourd'hui, ces consommateurs souhaitent des articles complets et actualisés sur tous les sujets imaginables, en un seul clic, dans leur langue maternelle.

Trop peu de gens ayant un accès à trop peu de données



Votre entreprise est confrontée au même problème avec ses données. Même si vous disposez des meilleurs experts, vous n'avez pas suffisamment de ressources pour mettre l'ensemble de ces données à la disposition de tous ceux qui en ont besoin, aussi rapidement qu'ils le souhaitent. Vous n'êtes pas non plus en mesure de répondre aux besoins croissants en types de données nouveaux et variés des utilisateurs métiers.

En fin de compte, les gens trouveront d'autres façons de répondre à leurs besoins en données, comme le « shadow IT » ou la création d'autres entités. Les équipes IT qui ne peuvent pas se détacher de ce modèle centralisé perdront rapidement le contrôle, compromettant ainsi la rapidité, l'exactitude et la sécurité.

L'accès aux données est étroitement contrôlé. Le modèle encyclopédique n'a pas su s'adapter à l'ère du big data, dans laquelle de nombreuses personnes exigent un accès immédiat aux bonnes données provenant de toutes les sources imaginables.

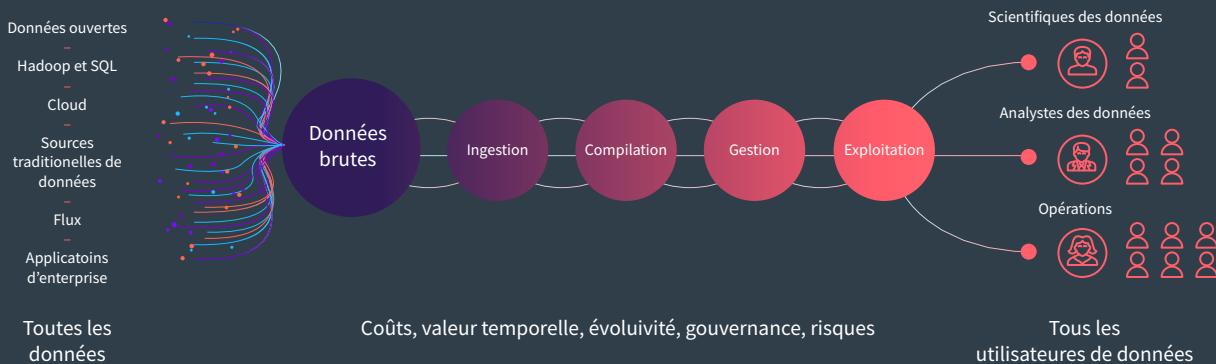
Le far west des données : la lutte pour contrôler la prolifération des données

L'avènement du big data a coïncidé avec l'essor d'une approche de la gestion des données beaucoup plus agile : le lac de données (« data lake »). Si l'approche décrite à la section précédente consistait à débuter par une modélisation des données avant de s'attaquer aux données selon une approche descendante, le data lake en est l'exact contraire. Tout commence par les données brutes. Ces données peuvent être ingérées avec des coûts de mise en œuvre initiaux minimes, en général dans des systèmes de fichiers de base et à bas coût. Pas besoin de s'embêter avec la structure des fichiers lors de l'arrivée des données. Il se peut même que vous ne sachiez pas ce qu'elle contient. Par la suite, vous pourrez créer une structure pour ces données (une étape que les spécialistes appellent « schema on read »), mais également mettre en place des contrôles de la qualité des données, des règles de sécurité, des politiques, des vérifications, etc.

Ce modèle plus agile présente plusieurs avantages par rapport au modèle centralisé. Il est étendu à différentes sources de données et différents cas d'usage. Et il touche différents publics ; seuls les experts en données peuvent toutefois accéder aux données brutes, les non-techniciens auront besoin de données structurées en lien avec leur métier pour pouvoir les exploiter.

C'est pourquoi les data lakes commencent généralement par une approche basée sur un laboratoire de données ciblant quelques data scientists experts en données. Grâce à une infrastructure cloud et aux big data, il est possible d'accélérer considérablement le processus d'ingestion des données brutes. L'utilisation du « schema on read » permet aux data scientists de transformer les données en données intelligentes de façon autonome.

Difficulté à contrôler la prolifération des données



Mais ce n'est pas tout. L'étape suivante consiste à partager ces données avec un plus large public qui a besoin de données plus structurées. La plupart des entreprises créent une nouvelle couche de données destinée à l'analytique, ciblant ainsi la communauté des analystes. Comme vous ciblez un public plus large avec des rôles différents, il devient alors évident que vous devez renforcer la gouvernance et le contrôle de la qualité des données.

Une fois que vous y êtes parvenu, l'étape suivante consiste à délivrer les informations à l'ensemble de l'entreprise. Par exemple, vous souhaitez peut-être recommander des produits à vos clients en intégrant un algorithme de machine learning dans vos applications de front office, ou monétiser certaines de vos données auprès de tiers. Là encore, l'élaboration d'une autre couche de gouvernance est une condition préalable.

Ce modèle plus agile présente plusieurs avantages par rapport au précédent. Il est étendu à différents cas d'usage, sources de données et publics. Les données brutes peuvent être ingérées au fur et à mesure de leur arrivée, avec des coûts de mise en œuvre initiaux minimes, et les modifications sont simples à appliquer.

Cette approche pose toutefois un défi de taille : la gouvernance des données n'a pas été pensée dans le même temps, mais après coup, lorsque les données ont dû être étendues à de nouveaux publics et cas d'usage.

Nous n'avons pas envisagé la gouvernance des données en même temps que ce modèle plus agile, mais plutôt après coup.

Revenons-en à notre métaphore du Web et étudions les défis auxquels certains réseaux sociaux, tels que Facebook, sont actuellement confrontés. Dans leur modèle initial, ces réseaux fournissaient uniquement la plate-forme, pas le contenu. Ceci a permis la création de communautés autogérées et, pendant un temps, ce modèle semblait

parfaitement adapté. En effet, leur plate-forme pouvait ingérer un volume de contenu et un nombre d'utilisateurs et de communautés illimités.

Puis vinrent les « fake news », les contenus violents et l'utilisation malveillante de la plate-forme. Comme n'importe qui peut publier n'importe quoi, sans aucune vérification, il est désormais très difficile de mettre en place un contrôle. De nouvelles réglementations en matière de confidentialité des données sont apparues et Facebook est aujourd'hui dans le collimateur des gouvernements et régulateurs, mais aussi de ses utilisateurs. Certains de ses plus grands annonceurs, comme Unilever, ont d'ores et déjà menacé publiquement de boycotter Facebook et Google si ces géants du Web ne contrôlent pas efficacement les contenus extrémistes et illégaux. En raison de ces pressions, la confiance et la sécurité sont devenues pour Facebook un thème central ces derniers mois et les frais associés sont en hausse. Facebook embauche 10 000 personnes supplémentaires affectées à la problématique de la fiabilité et de la sécurité, et prévoit de doubler ses effectifs à l'avenir.

Les enseignements de cette situation ? Une entreprise n'a aucune chance de réussir à l'ère du numérique sans définir un cadre. La réussite de votre transformation numérique dépend de la gouvernance des données et, lorsque vous vous en rendrez compte, il sera peut-être déjà trop tard. Sans la gouvernance, votre lac de données se transformera rapidement en marécage. Et, à ce stade, votre entreprise pourrait devoir faire des efforts considérables pour se désencombrer et tirer à nouveau parti en toute sécurité de ses données.

Fournir des données en temps réel est le Saint-Graal de toute entreprise mais on ne transige pas avec la gouvernance des données.

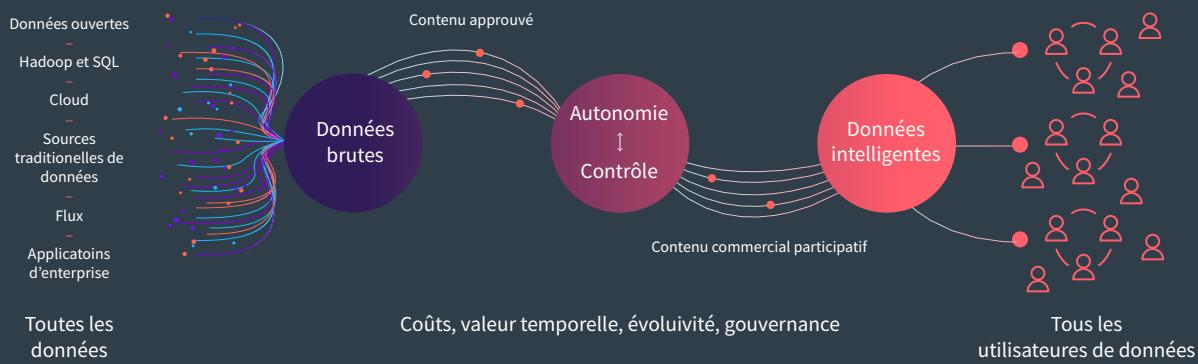
Mettre en place une gouvernance collaborative à l'ère du numérique

Le juste équilibre entre gouvernance ascendante et descendante

Le second modèle ne permet pas de prendre le contrôle des données au fur et à mesure de leur entrée dans vos systèmes. Mais, dans le même temps, il faut reconnaître que ces données émanent de sources, toujours plus nombreuses et d'un nombre croissant de collaborateurs de différents départements de l'entreprise. En mettant en place d'emblée une approche plus collaborative de la gouvernance, vos utilisateurs métiers les plus experts peuvent devenir des fournisseurs de contenu et des data curators. Il est essentiel pour cette approche de travailler en équipe dès le départ. Dans le cas contraire, vous risquez d'être submergé par la quantité de travail qu'implique la confirmation de la fiabilité de vos données.

Penchons-nous une fois encore sur les enseignements tirés de l'ère du Web. Wikipedia s'est imposé comme l'un des 5 sites les plus visités à travers le monde. Il héberge plus de 5 millions d'articles, mais seuls 1 194 administrateurs gèrent ses différentes pages. Mais n'importe qui peut apporter une contribution et le site compte 130 000 auteurs réguliers. Pour faire face à cette situation, Wikipedia a instauré des principes clairement définis pour la compilation des contenus collaboratifs. Le site a ainsi démontré sa capacité à s'adapter et à fournir un contenu qui bénéficie d'un niveau de fiabilité correct.

Développer la confiance et la portée par la collaboration

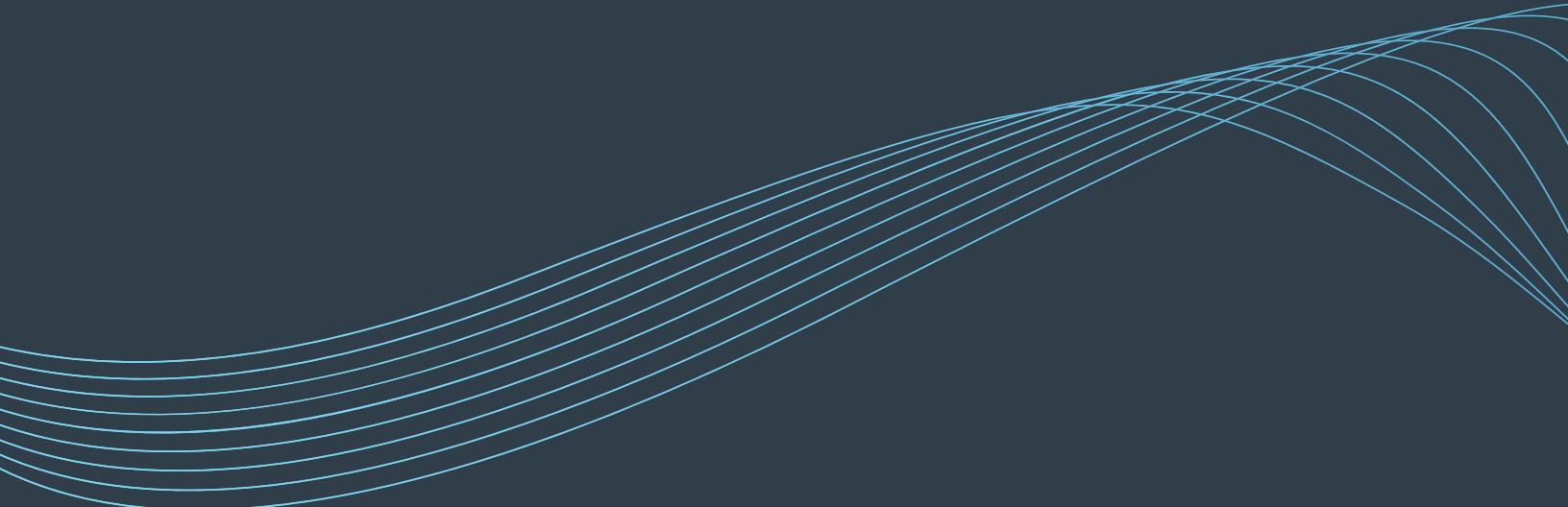


En adoptant une approche semblable à celle de Wikipedia dans laquelle n’importe qui peut collaborer à la compilation des données tant que les normes sont respectées, vous aiderez votre entreprise à transformer ses données brutes en informations fiables, documentées et prêtées à être partagées. Grâce à des outils en libre-service intelligents et orientés workflows qui intègrent des contrôles de la qualité des données, vous pouvez mettre en place un système fiable.

Ce modèle peut venir en complément de l’approche descendante, au lieu de la remplacer. En effet, certains processus lourdement réglementés, tels que l’agrégation des données sur les risques dans les services financiers, et certaines données spécifiques telles que les coordonnées de cartes de crédit, exigent une attention toute particulière et, dans ce cas, l’approche ascendante ne s’appliquera pas. Entre autres responsabilités, l’équipe chargée de la gouvernance des données doit définir le modèle à appliquer à l’entreprise.

Dans le chapitre suivant, nous étudierons comment mettre en oeuvre le modèle collaboratif et gouverné selon une approche en trois étapes.

Un modèle de gouvernance collaborative de type Wikipedia permet à vos utilisateurs métiers les plus experts de devenir fournisseurs de contenu et « data curators ».



Chapitre 3 :

Trois étapes pour fournir rapidement des données fiables

Mettre en œuvre une gouvernance maîtrisée et collaborative

Bien que la technologie puisse permettre de résoudre le problème, les entreprises doivent mettre en place un mode d'organisation de leurs données à grande échelle. Comme nous l'avons vu au chapitre précédent, il est capital de réinventer la gouvernance traditionnelle des données pour faire face à leur prolifération : selon Gartner, « d'ici 2022, seuls 20 % des entreprises investissant dans l'information parviendront à adapter la gouvernance au numérique ». Compte tenu du nombre important d'entreprises envahies par les données, ce pourcentage est à l'évidence bien trop faible.

Une gouvernance moderne des données ne vise pas simplement à réduire les risques liés aux données et à créer une « data policy ». Son but consiste également à optimiser l'utilisation des données, raison pour laquelle les approches de gouvernance traditionnelle et rigoureuse des données sont insuffisantes.

Il est à la fois nécessaire et possible d'adopter une approche ascendante plus agile. Cette stratégie associe tout d'abord les données brutes à leur contexte métier de façon à ce qu'elles soient pertinentes, instaure un contrôle de leur qualité et de leur sécurité et les organise de manière rigoureuse en vue d'une consommation massive. C'est le moment pour votre entreprise d'adopter cette approche, car de plus en plus de gens prennent conscience de la valeur des données et des avantages d'une gestion adéquate. Et, suite aux scandales et aux fuites de données qui ont fait la une des journaux et face à la multiplication des nouvelles réglementations sur la protection des données, ils en comprennent maintenant aussi les enjeux.

De nouvelles plates-formes de données dynamisent cette nouvelle discipline, qui tire parti de technologies importantes (comme la reconnaissance de structure, le catalogage des données, le lignage des données et le machine learning) pour organiser les données à grande échelle et transformer la gouvernance des données en un sport d'équipe. Elles permettent une collaboration à l'échelle de l'entreprise en matière de propriété, de compilation, de remédiation et de réutilisation des données.

Dans ce chapitre, nous aborderons une méthodologie en trois étapes afin de fiabiliser, fédérer et diffuser les données au sein de votre entreprise. À l'aide d'exemples, nous illustrerons également la façon dont les plates-formes d'intégration de données peuvent faciliter la mise en œuvre de cette approche au moyen d'outils et de fonctionnalités adéquats.

À retenir :

Assurez-vous que vous pouvez découvrir et nettoyer vos données dès leur entrée dans votre environnement de données. Organisez et donnez les moyens d'agir, en créant par là même un point unique de données fiables où vous définissez la responsabilité des données et donnez les moyens d'agir à ceux qui doivent les documenter, les protéger et les diffuser largement, en équipe. Maintenant que vous avez le contrôle, vous pouvez automatiser et déployer, de façon à pouvoir fournir des informations à grande échelle à un large éventail de consommateurs de données et d'applications.

1^{re} étape : découvrez et nettoyez vos données

Les solutions de stockage de données étant devenues de plus en plus abordables et accessibles au cours des dernières années, de grands référentiels de données tels que les data lakes ont vu le jour. Les équipes doivent donc gérer un nombre croissant d'ensembles de données connus et inconnus de tous genres. Ces ensembles de données, espérons-le, viendront enrichir votre data lake. Malheureusement, c'est parfois le contraire. Dans ce cas, l'ensemble de l'entreprise accumule les retards dans le traitement des données. Si un logiciel moderne peut ingérer ces données en quelques secondes seulement, les équipes IT peuvent mettre des semaines à publier de nouvelles sources de données dans des entrepôts de données ou des data lakes.

Parallèlement, il se peut que les consommateurs de données ne sachent même pas que les données qu'ils cherchent sont à leur disposition. Les analystes ou les data scientists mettent des heures à trouver, comprendre et contextualiser toutes les données. IDC a montré que les professionnels des données ne consacrent que **19 %** de leur temps à analyser les informations et à produire des résultats utiles contre 37 % à préparer des données et 24 % à les protéger. IDC estime que 12 heures sont ainsi perdues chaque semaine. Et la gouvernance devient un vrai casse-tête, car, si les collaborateurs de l'entreprise ne trouvent pas les données qu'ils recherchent, ils les recréeront. Et appliqueront ensuite leurs propres règles aux sources de données qu'ils viennent de créer, constituant ainsi à terme plusieurs référentiels.

La difficulté consiste à surmonter ces obstacles et à apporter clarté, transparence et accessibilité à vos ressources de données. Quel que soit l'emplacement de ces données (dans applications d'entreprise comme Salesforce.com, Microsoft Dynamics ou SAP, un entrepôt de données traditionnel ou un lac de données cloud), commencez par passer vos données au crible pour avoir une vue globale des sources et flux de données qui entrent et sortent de votre entreprise.

À retenir :

N'opérez pas à l'aveugle avec vos données. Dans vos premiers pas vers votre stratégie de données, vous devez savoir ce que vos données contiennent. Pour cela, vous avez besoin d'outils et de méthodologies pour renforcer votre stratégie orientée données.

En premier lieu, vous devez connaître et comprendre vos données

Lors du traitement de vos données, il est vital de commencer à explorer les différentes sources de données que vous souhaitez gérer. Vous devez identifier les schémas et étudier les ensembles de données et la structure des données provenant des différentes sources à votre disposition.

Par le passé, cette activité était réalisée manuellement par des spécialistes, à l'aide d'outils classiques de profilage des données. Les ensembles de données connus y étant traités un par un, cette approche ne marche plus. La prolifération des données dans l'ère du numérique nécessite une méthode plus automatique et systématique. D'où l'utilité d'outils modernes de catalogage des données tels que Talend Data Catalog. Il vous permettra de programmer les processus de découverte des données qui exploreront votre data lake ou tout autre environnement de données et inspecteront avec intelligence les données sous-jacentes de façon à ce que vous puissiez les comprendre, les documenter et prendre des mesures en fonction du contenu de vos ensembles de données.

Le profilage automatique à la portée de tous avec Data Catalog

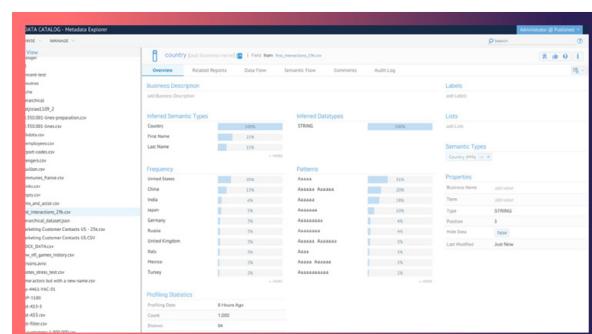
Grâce aux capacités de profilage automatique de [Talend Data Catalog](#), les non-spécialistes peuvent contrôler aisément les données de votre entreprise. Avec Talend Data Catalog, vous profitez de fonctions de découverte automatique et de documentation intelligente des ensembles de données de votre data lake. Vous pouvez utiliser des fonctionnalités de profilage et d'échantillonnage pour évaluer vos données d'un simple coup d'œil. Avec des ensembles de données fiables avec profilage automatique, vous disposez d'indicateurs performants et visuels permettant aux utilisateurs de trouver facilement les données pertinentes en quelques clics.

Talend Data Catalog peut établir automatiquement les relations entre les ensembles de données et les associer à un glossaire métiers. En somme, cela permet à une

entreprise d'automatiser l'inventaire de ses données et d'utiliser une sémantique intelligente pour le profilage automatique, la découverte des relations et la catégorisation grâce à un flux sémantique intégré.

Les avantages sont doubles : les fournisseurs et les propriétaires des données bénéficient d'une vue d'ensemble de leurs données et peuvent prendre les mesures qui s'imposent. Par exemple, les éléments de données critiques méritant une attention particulière, tels que les données personnelles et la conformité avec la législation en matière de protection des données, sont automatiquement catégorisés et mis en évidence. Le catalogue de données mettra également en lumière les éventuels problèmes de qualité des données nécessitant la mise en place de mesures correctives.

Les consommateurs de données peuvent désormais voir le contenu des données avant de les utiliser, au travers d'échantillons ou d'indications signalant, par exemple, qu'une certaine colonne contient des numéros de téléphone, des numéros de compte ou des adresses e-mail.



» Figure 1 : Profilage des données avec Talend Data Catalog

Allez encore plus loin avec le profilage des données

Le profilage des données est une technologie qui vous permettra une découverte en profondeur de vos ensembles de données et une évaluation précise de vos nombreuses sources de données selon les six dimensions de la qualité des données. Vous pourrez ainsi repérer plus facilement si des données sont erronées, incohérentes ou incomplètes et de quelle manière.

Imaginez qu'un médecin fasse passer un examen à un patient pour évaluer son état de santé. Personne ne veut se faire opérer sans subir au préalable un examen approfondi. La logique est la même pour le profilage des données. Vous devez comprendre vos données avant de les réparer. Les formats des données que vous êtes amené à traiter sont souvent inexploitables, non structurés ou masqués. Vous devez donc établir un diagnostic précis pour mieux comprendre les problèmes avant d'y remédier. Vous, votre équipe et l'ensemble de votre entreprise gagnerez ainsi du temps, car vous aurez d'ores et déjà dressé un état des lieux approfondi.

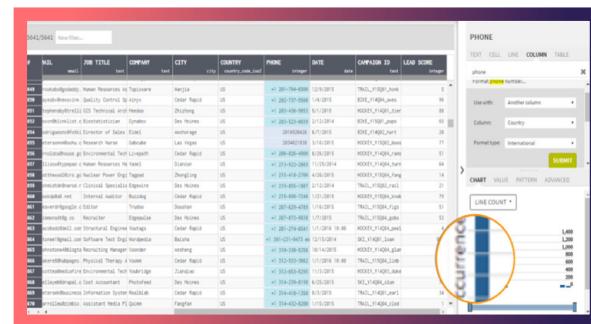
De la même façon qu'un médecin généraliste et un spécialiste jouent des rôles différents, mais cruciaux dans le diagnostic médical et l'établissent en s'appuyant sur des approches et des outils légèrement différents, les techniques de profilage des données techniques s'appliquent à différents rôles et nécessitent des outils à part.

80 % des entreprises interrogées déclarent qu'elles ont été concernées par le RGPD ou une autre loi sur la protection des données.

Offrez aux utilisateurs avancés un profilage en libre-service

Dans de nombreux cas, ceux qui connaissent le mieux les données ne sont pas les experts en données. Prenez, par exemple, les coordonnées de vos clients : les administrateurs des ventes, commerciaux et responsables marketing terrain sont plus au fait des problèmes de qualité des données que l'équipe IT centrale. Et ils pâtissent le plus de ces problèmes, car ils ont une incidence sur l'efficacité avec laquelle ils peuvent faire leur travail au quotidien. Pour nettoyer les données de Salesforce, vous souhaiterez peut-être évaluer leur qualité en dépliant certaines activités de profilage à ces utilisateurs métiers.

Bien évidemment, il n'est pas question de leur demander de devenir des spécialistes de la qualité des données. Un nouveau type d'outil intelligent permettant de dissimuler la complexité technique et d'offrir une expérience utilisateur simple, rapide et visuelle pour accélérer le profilage de leurs ensembles de données de prédilection est nécessaire pour cela. Grâce à des outils tels que [Talend Data Preparation](#), vous disposerez de fonctionnalités de profilage intégrées à la fois simples et puissantes pour explorer les ensembles de données et évaluer leur qualité au moyen d'indicateurs, de tendances et de schémas.



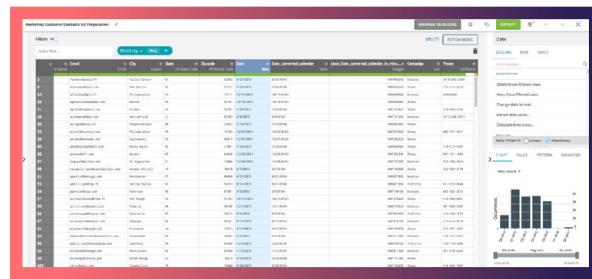
» Figure 2 : Profilage des données pour les utilisateurs avancés avec Talend Data Preparation

Instaurez la confiance dans vos données grâce au profilage avancé

Si le profilage automatique des données grâce à un catalogue de données et un profilage en libre-service est utilisé pour la gouvernance des données ascendante, une approche descendante peut nécessiter un examen plus approfondi des données. Prenons, par exemple, le cas de l'agrégation et du reporting des données sur les risques. Ils sont définis par des principes formels et des réglementations connexes. Ils reposent notamment sur le principe suivant : « Les contrôleurs ou « supervisors » attendent des banques qu'elles mesurent et contrôlent l'exactitude des données et mettent en place des canaux de remontée d'information et des plans d'action appropriés afin de remédier à la mauvaise qualité des données ».

Ce type d'approche formelle, lorsqu'elle ne satisfait pas aux normes, nécessite l'intervention d'une équipe de DevOps (ingénieurs des données, spécialistes de la qualité des données ou développeurs informatiques). Grâce à des outils tels que Talend Data Quality dans Talend Studio, ils commenceront par accéder aux sources de données pour analyser leur structure (catalogues, schémas et tableaux) et stockeront la description de leurs métadonnées dans un référentiel de métadonnées. Ils définiront ensuite les analyses de qualité des données disponibles, y compris les analyses des bases de données, des contenus, des colonnes, des tableaux, des redondances et des corrélations. Ces analyses exécuteront des processus de profilage des données pour définir le contenu, la structure et la qualité de structures de données très complexes.

Enfin, cette découverte approfondie établira un « indice de fiabilité » qui sera calculé et rapporté, et qui fera l'objet d'un suivi régulier automatisé. Cela permettra également le déclenchement d'alertes lorsque cet indice de fiabilité est inférieur à un certain seuil.



» Figure 3 : Profilage des données avancé pour les data engineers

À retenir :

Gardez à l'esprit que votre stratégie de données doit avant tout débuter par la découverte des données. En ne procédant pas au profilage de vos données, vous risquez de mettre en péril l'ensemble de cette stratégie. Vous devez analyser le terrain pour vous assurer que vous construisez votre entrepôt de données sur des fondations solides.

Prenons, par exemple, une initiative de confidentialité des données qui ne permet pas de déterminer si un nouveau dataset contenant des données personnelles a pénétré dans votre environnement de données. L'ensemble de votre programme de conformité pourrait être mis en péril par ces nouvelles données que ne vous n'avez pas pu identifier.

62 % des utilisateurs reconnaissent avoir accès à des données qu'ils ne devraient pas pouvoir consulter.

Intégrité des données garantie dès le départ

Il est primordial d'intégrer des contrôles qualitatifs dans votre chaîne de données pour garantir la réussite de votre initiative de gouvernance.

Imaginons, par exemple, que vous souhaitiez lancer une campagne visant à contacter des clients concernant la facturation et les paiements et que vous disposez principalement d'adresses e-mail et postales pour les contacter. Il est essentiel que ces données soient exactes et uniformes pour joindre tous ces clients, faute de quoi vous risquez de perdre beaucoup d'argent ou de passer à côté d'opportunités commerciales en raison de données manquantes ou incohérentes.

Les problèmes d'intégrité des données se sont multipliés ces dernières années. Comme nous l'avons vu précédemment, l'augmentation des sources et du volume de données entraîne une hausse du nombre de professionnels désireux de les exploiter. Cette prolifération de données dans un nombre croissant de clouds et de canaux numériques et la multiplication d'acteurs très différents augmentent la vulnérabilité de l'entreprise, exposée à des fuites, des violations des données, ainsi qu'à des informations erronées reposant sur des données indésirables et incohérentes. À titre d'exemple, **62 %** des utilisateurs reconnaissent avoir accès à des données qu'ils ne devraient pas pouvoir consulter. L'intégrité est devenue une question d'autant plus essentielle que la mise en œuvre de nouvelles réglementations en matière de gouvernance impacte de manière concrète les entreprises. Par exemple, l'amende pour le non-respect du Règlement général sur la protection des données (RGPD) de l'Union européenne s'élève à 4 % du chiffre d'affaires mondial de l'entreprise.

Pourtant, il est possible de résoudre ce problème. Les entreprises doivent assurer l'exactitude et la disponibilité des données pour tous ceux qui en ont besoin.

Elles ne doivent pas se reposer uniquement sur une petite équipe IT ou une poignée de collaborateurs spécialistes des données pour y parvenir. Tous, du département IT aux data scientists, en passant par les intégrateurs d'applications et les analystes, doivent pouvoir tirer de précieuses informations des données de bonne qualité.

Orchestrez l'intégrité des données dans les pipelines

La qualité des données implique de préparer les données afin qu'elles répondent aux besoins spécifiques des utilisateurs métiers. L'exactitude, l'exhaustivité, la cohérence, la ponctualité, la singularité et la validité constituent les principales mesures de la qualité des données.

Mais la qualité des données ne peut pas être un processus isolé. Pour le mener à bien et fournir des données fiables, vous devez effectuer les opérations de qualité des données dès le début et nativement, à partir des sources de données, en parallèle du cycle de vie des données, afin que les opérateurs de données, les utilisateurs ou les applications puissent au final exploiter des données fiables.

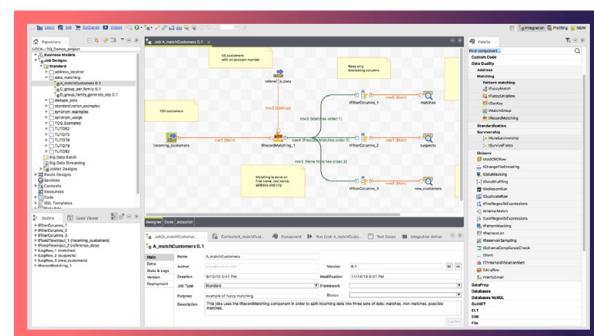
Ainsi, avant d'établir votre point unique de données fiables et de les diffuser, il est crucial de soumettre les sources de données ingérées à des contrôles de qualité des données et leur appliquer des mesures correctives. Talend Data Quality génère pour cela un code natif afin d'effectuer ces contrôles au bon endroit (sur site, dans un cluster de big data ou dans le cloud) et au bon moment (sur des données statiques ou de streaming).

Vous devrez profiler, nettoyer et normaliser vos données ainsi que surveiller l'évolution de leur qualité sur la durée, quels que soient le format ou le volume considérés.

C'est pourquoi pour garantir leur qualité, plutôt que des solutions monofonctionnelles isolées, vous avez besoin d'une plate-forme répandue offrant un large éventail de contrôles de la qualité des données non seulement pour le nettoyage et la normalisation des sources de données, mais aussi pour déléguer certaines tâches aux experts grâce à des outils intégrés en libre-service.

Pour la gouvernance, la qualité des données est essentielle pour nous — tout comme le fait de connaître leur provenance.

Dirigeant, entreprise d'assurance maladie

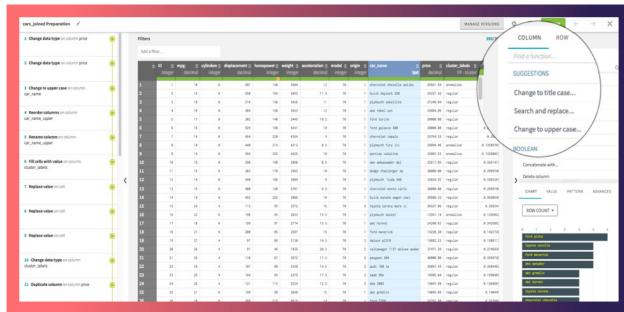


» Figure 4 : Talend Data Preparation orchestre les données avec l'intégrité à l'aide des pipelines

Déléguez le nettoyage des données grâce au libre-service

Comme nous avons pu le voir, les données ne relèvent plus de la seule responsabilité d'une organisation centrale. Un cadre de gouvernance des données performant implique d'établir des responsabilités, puis de déléguer cette autorité de façon pertinente. Par exemple, un responsable de la protection des données dans une organisation centrale pourra souhaiter déléguer certaines tâches à des data stewards ou des utilisateurs métiers : un ingénieur commercial peut en effet être mieux placé pour s'assurer que les coordonnées de ses clients sont exactes et actualisées. Le gestionnaire de campagnes doit s'assurer qu'un mécanisme de consentement a été mis en place et intégré à sa base de données marketing.

Afin de faciliter ce type de délégation, les entreprises doivent doter les différents départements d'applications en libre-service basées sur des workflows, telles que Talend Data Preparation. Ils jouiront ainsi d'une autonomie accrue sans risquer de compromettre les données.



» Figure 5 : Nettoyage des données en libre-service avec Talend Data Preparation

À retenir :

Les responsables des données ne peuvent pas maîtriser la transformation numérique – et ses répercussions – avec les rôles et organisations centralisés d'hier qui génèrent des points de ralentissement et des décalages. La gouvernance des données est un « sport d'équipe » dans lequel plusieurs fonctions et départements doivent jouer en collectif.

Améliorez la qualité de vos données grâce au cloud

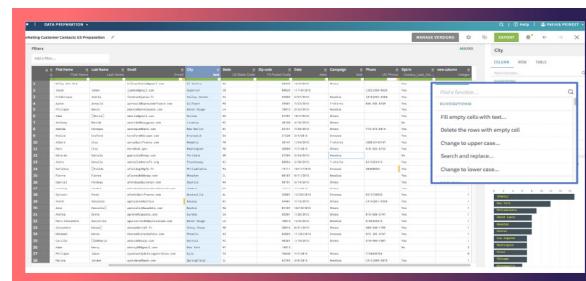
Le cloud repousse considérablement les limites des données. Les divers départements utilisent leurs propres applications, et les produits, les personnes et les ressources créent leurs propres pipelines de données via le Web et l'Internet des objets. Les divers acteurs de l'entreprise et les fournisseurs de données peuvent échanger des données en toute transparence.

Le libre-service permet d'adapter les normes de qualité aux besoins. De nombreuses études ont montré que, lorsque les données fiables ne sont pas fournies en libre-service, les analystes et les data scientists passent 80 % de leur temps à les nettoyer et à les rendre exploitables. Moins on y consacre de temps et d'efforts, plus les coûts sont réduits. Et, par conséquent, ces données génèrent plus de valeur ajoutée et d'informations.

Avec des applications en libre-service comme Talend Data Preparation qui répond à ce problème, tous les collaborateurs peuvent accéder à un ensemble de données, puis nettoyer, standardiser, transformer ou enrichir les données. Talend Data Preparation étant simple à utiliser, cette solution résout un problème central des entreprises où de nombreux collaborateurs passent un temps considérable à traiter des données dans des feuilles Excel ou à attendre que des collègues s'en chargent pour

La préparation des données ne consiste pas seulement à rendre les départements plus autonomes dans la gestion des données ; c'est un élément clé de la qualité et de l'intégration des données. Cette étape améliore la productivité lors de la gestion des données et permet également de garder une trace des actions menées par chacun sur les données. Lorsque ces actions contribuent à renforcer la fiabilité des données, elles peuvent être déployées et intégrées dans les pipelines de données de façon à ce qu'elles profitent à chacun. Outre l'amélioration de la productivité individuelle, le véritable intérêt de ces applications collaboratives en libre-service est de favoriser la collaboration entre les fonctions métiers et le département IT.

Au terme de la première des 3 étapes visant à fournir des données fiables, les sources de données ont été identifiées et documentées. Des mesures ont été prises concernant les sources dont la qualité n'est pas au rendez-vous.



» Figure 6 : Accès en libre-service avec Talend Data Preparation

À retenir :

Avant de choisir une plate-forme de gouvernance des données, vous devez déterminer si elle vous permettra de déléguer les opérations liées à la qualité des données en mode libre-service à des utilisateurs métiers tout en gardant le contrôle. Ce point est essentiel pour adapter et mutualiser rapidement les efforts déployés pour nettoyer les données en temps réel. Il serait risqué de ne rien faire et de laisser vos collaborateurs les préparer et les nettoyer seuls, et consacrer, ce faisant, un temps considérable à des tâches répétitives sur des sources de données non maîtrisées.

2^e étape : organisez les données fiables et donnez les moyens d'agir

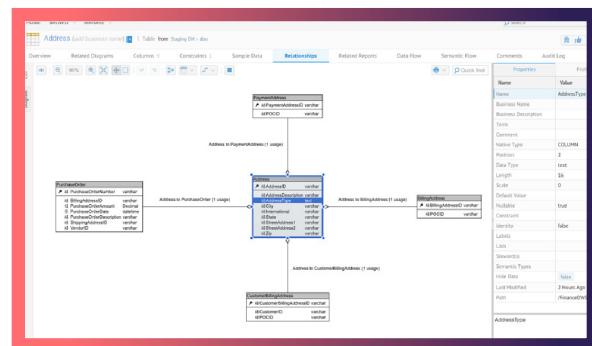
Organisez les données avec intégrité en établissant un point unique de données fiables

Au terme de la première étape, vous avez pu identifier les ressources de données entrantes, les documenter et vérifier leur fiabilité. Il est temps à présent de les organiser pour une utilisation intensive par le réseau étendu d'utilisateurs de l'entreprise.

Pour commencer, mettez en place un point unique de données fiables. Autrement dit, réunissez tous les ensembles de données en un seul point de contrôle qui sera la clé de voûte de votre cadre de gouvernance des données. Vous devrez ensuite sélectionner les ensembles de données identifiés, attribuer les rôles et les responsabilités directement via votre point de contrôle unique afin que votre gouvernance soit opérationnelle dès le début.

Le catalogage des données présente l'avantage de pouvoir regrouper toutes les données fiables en un emplacement unique auquel les utilisateurs peuvent accéder. Ils peuvent ainsi les utiliser immédiatement, les protéger, les agréger et autoriser diverses personnes et applications à en tirer parti.

Autre avantage : en centralisant les données fiables dans un environnement où elles peuvent être partagées, votre entreprise économise du temps et des ressources.



» Figure 7 : Mise en place d'un espace centralisé pour les données fiables avec Talend Data Catalog.

À retenir :

Selon « Magic Quadrant for Business Intelligence and Analytics Platforms », Gartner, 2017 : « [D'ici 2020](#), les entreprises qui offrent à leurs utilisateurs un accès à un catalogue organisé de données internes et externes multiplieront par deux la valeur commerciale de leurs investissements analytiques par rapport aux autres. »

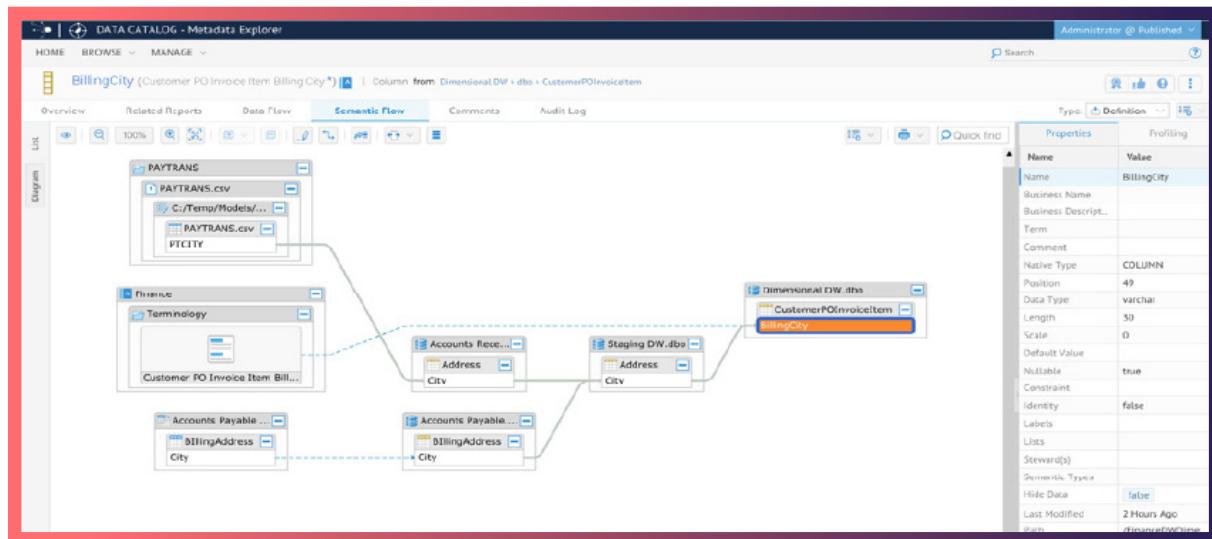
Définissez vos données au sein d'un glossaire métiers et rendez-les accessibles grâce au lignage des données

Dans un catalogue de données, un glossaire métiers permet de définir des ensembles de termes et les associer à des catégories et sous-catégories. La constitution d'un glossaire métiers peut être extrêmement simple : utilisez un modèle de données existant bien documenté, importez les termes et les définitions d'autres sources (par exemple, CSV, Microsoft Excel) ou rédigez-les de façon interactive via l'interface utilisateur au cours de la catégorisation des objets. Une fois publié, le glossaire est en théorie accessible par toutes les personnes disposant des autorisations nécessaires par le biais d'une interface de recherche (voir figure ci-dessous).

Lorsque vous laissez vos collaborateurs accéder à votre catalogue, votre ensemble de données prend vie, car vous permettez à des personnes autorisées de modifier, valider et enrichir les données au sein d'ensembles de données. Vous économiserez beaucoup de temps et de ressources en permettant que ces opérations soient réalisées par le biais d'un catalogue de données.

Encore plus important, le lignage des données vous offrira une vue d'ensemble vous permettant de suivre vos flux de données depuis leur source jusqu'à leur destination finale.

Imaginez que vous constatiez que des données incohérentes ont été créées dans vos systèmes de données et reprises dans l'un de vos ensembles de données et que l'on vous demande de les expliquer, les identifier et les corriger. Le lignage des données vous sera d'une aide précieuse pour cela. Il vous permettra de repérer le problème et son emplacement, et garantira l'exactitude permanente de vos données. En outre, en cas d'ajout de nouveaux ensembles de données à votre data lake, un lignage des données vous aidera à identifier très rapidement ces nouvelles sources.



» Figure 8 : Constituer un glossaire métiers avec Talend Data Catalog.

Le lignage des données vous permettra de répondre rapidement aux demandes d'accès aux pistes d'audit des autorités compétentes : parmi les exemples concrets, on peut citer les réglementations en matière de respect de la vie privée (RGDP ou CCPA) ou les réglementations relatives aux risques d'atteinte aux données pour les services financiers (BCBS 239). Ces dernières encouragent la création d'un inventaire des données toujours exact capable de suivre la provenance des données et les activités annexes de traitement des données appliquées à certaines.

Les questions récurrentes sont, notamment : lorsqu'un rapport de gestion fait état d'une erreur, d'où provient-elle ? Quand l'erreur s'est-elle produite ? Qui en est responsable ? Comment pouvez-vous la corriger ? Une solution de gestion des métadonnées intégrant le lignage des données apportera une réponse à toutes ces questions. Le lignage vous permettra d'avoir rapidement un aperçu général des vues de données et donc de détecter facilement le problème.

Il joue également un rôle important dans la gestion des changements. Par exemple, imaginons qu'il faille mettre à jour un élément de données dans une chaîne de données, quel en sera l'impact en amont de la chaîne de données ? Quels seraient les autres éléments de données également affectés ? Répondre à ces questions prendrait des semaines. Alors qu'avec des principes et outils adéquats de catalogage et de lignage des données, quelques clics pourraient suffire.

« Les logiciels de data intelligence sont essentiels pour la gouvernance des données et les programmes correspondants, car ils dotent les entreprises des connaissances dont elles ont besoin pour fournir des données pertinentes à la bonne ressource au meilleur moment. Le lignage des données est un élément clé de ces solutions d'analyse intelligente : il offre des informations supplémentaires sur les données et dégage encore plus de résultats et de valeur ajoutée pour les entreprises orientées données. » -- Stewart Bond, directeur de la recherche sur l'intégration et l'intégrité des données chez IDC.

Stewart Bond, director of Data Integration and Data Integrity research at IDC.

Identifiez les rôles et responsabilités

Une fois vos catégories de données ou « éléments de données critiques » (EDC) définis, vous pourrez vous faire une idée plus précise des sources présentes dans votre environnement de données.

Ceci vous permettra également de mieux définir les propriétaires des données : qui est responsable d'un domaine de données précis ? Qui est chargé de la consultation, l'accès, la modification et la compilation des ensembles de données ?

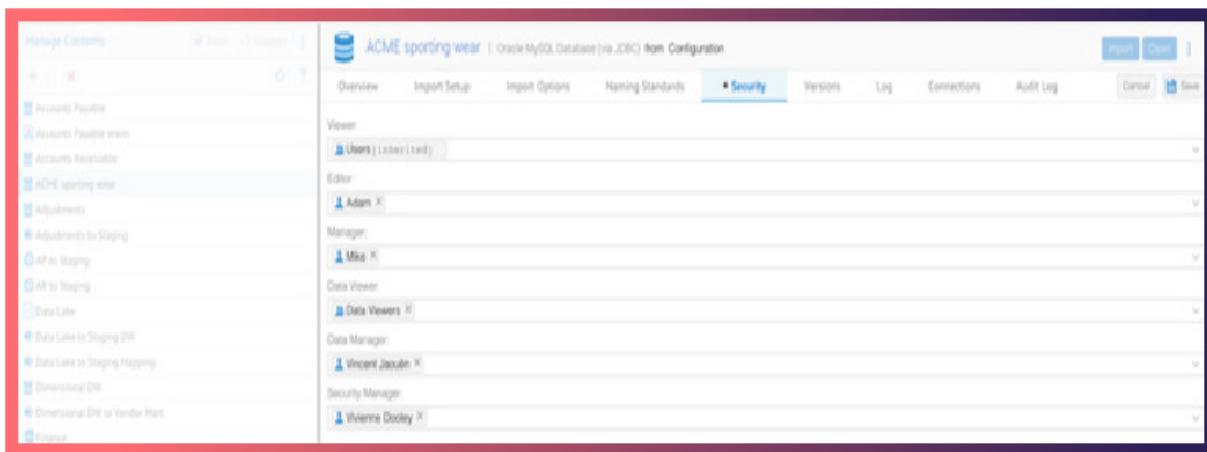
À ce stade, l'utilisation d'un modèle RACI* vous fera gagner du temps lors de la définition et l'attribution des rôles et des responsabilités des différents acteurs de votre entreprise.

L'étape suivante consiste à définir les propriétaires de données qui sont responsables en dernier ressort d'une ou plusieurs catégories et sous-catégories de données. Ces propriétaires de données seront chargés d'effectuer les opérations courantes menées sur les données ou de les déléguer à des data stewards. Ils identifieront les

ensembles de données et les EDC, et mettront en place des normes pour le recueil, l'utilisation et le masquage des données. Talend Data Catalog peut également inventorier les propriétaires et les stewards des différentes catégories et sous-catégories de données, puis leur attribuer des rôles et des workflows.

Par exemple, cette solution vous permet d'inventorier les propriétaires de données des catégories « client », « identité client », « facturation client », « coordonnées client » et « adresse de livraison client ».

* L'acronyme RACI est formé à partir des initiales (en anglais) des quatre responsabilités clés les plus couramment utilisées : Responsible (responsable), Accountable (autorité), Consulted (consulté) et Informed (informé). Le modèle RACI est un bon modèle de matrice d'attribution des qualités. Il est facile à comprendre et à utiliser. Il est particulièrement utile si votre gouvernance des données concerne plusieurs départements et divisions de votre entreprise.



» Figure 9 : Identification des rôles et responsabilités avec Talend Data Catalog

Donnez à vos collaborateurs les moyens de mener à bien la compilation et la remédiation des données

Le nombre de personnes qui se connectent à vos sources comme la confiance qu'ils accordent à vos données sont des mesures du succès de votre stratégie. Pour garantir ce succès, vous devez proposer des applications simples d'utilisation et prenant en compte les rôles de chacun, qui vous permettront non seulement de faire participer vos collaborateurs, mais aussi d'assurer leur responsabilisation quant à l'exactitude et la valeur des données.

Le nettoyage et la consolidation des données clients nous permettent de fournir le type d'expérience personnalisée que les clients méritent et attendent aujourd'hui.

Steve Brennan, vice-président de la stratégie et l'analytique chez Carhartt, Inc.

Donne à vos collaborateurs les moyens de procéder à la compilation des données

[Wikipedia](#) définit la compilation de données comme « l'organisation et l'intégration des données issues de sources variées. Cela inclut l'annotation, la publication et la présentation des données pour en garantir la validité au fil du temps. » Elle sera possible une fois que vous aurez mis en place un modèle RACI clair dans lequel vous déterminerez les personnes pouvant définir, modifier, valider et enrichir les données se trouvant dans vos systèmes.

Il est temps maintenant d'inciter vos collaborateurs à participer. La communication est essentielle.

Supposons que vous ayez décidé avec votre responsable exécutif d'annoncer officiellement le lancement de votre projet. Impliquez votre département Communication interne de façon à expliquer l'objectif général de ce projet.

Pourquoi est-il important de donner à vos collaborateurs les moyens de compiler des données ? La gouvernance des données ne consiste pas seulement à autoriser l'accès généralisé à des données fiables. Il s'agit également de mettre en avant la responsabilisation des dépositaires des données vis-à-vis du reste de l'entreprise de façon à ce qu'ils puissent enrichir les données fiables, les compiler, puis produire des informations précises et utiles à partir des pipelines de données.

Utilisez les mails et les systèmes de collaboration internes ou choisissez un nom ou un visuel percutant pour illustrer l'objectif de votre programme de données. Parallèlement à votre plan de communication, assurez-vous de toucher différents publics et départements utilisant des données dans leurs activités quotidiennes. Ainsi, votre programme de communication sensibilisera vos collaborateurs les plus au contact des clients, des services ou des produits : ce sont eux que vous voulez à vos côtés pour ce programme.

Donnez à vos collaborateurs les moyens de corriger les données

Souvent, les propriétaires de données se rendent compte qu'ils doivent endosser le rôle de chef d'orchestre plutôt que gérer eux-mêmes tout ce qui touche à leurs données. Le caractère collaboratif de la gestion des données paraît alors tout à fait fondamental. Vous devrez faire appel, de manière occasionnelle ou régulière, aux spécialistes en la matière pour la certification, l'arbitrage, la résolution ou le rapprochement des données. Une application telle que [Talend Data Stewardship](#), permet aux data stewards de concevoir, d'orchestrer et de lancer des « campagnes de stewardship » appelant certains collaborateurs identifiés à apporter une contribution essentielle qui viendra enrichir dynamiquement vos données.

Il est donc possible de devenir data steward à tout moment et de prendre part à la chaîne de valeur des données. Ces data stewards corrigent et valideront rapidement les données incohérentes au sein d'une application conviviale, entièrement déployée par le responsable de la campagne « stewardship ». operationalized by the steward campaign manager.

The screenshot shows the Talend Data Stewardship application interface. At the top, there's a navigation bar with 'DATA STEWARDSHIP' and other options like 'Help' and 'user1@user1.last'. Below the header, it says 'Campaign: Demo CRM data deduplication - user1@scorreia-dw.com' and 'State: New'. A 'VALIDATE CHOICES' button is also visible.

The main area is a table titled '11/11' with columns for ID, FIRST_NAME*, LAST_NAME*, GENDER*, AGE*, OCCUPATION*, COMPANY*, ADDRESS*, CITY*, and STATE*. The table contains 11 rows of data, each with a small circular icon and a red 'X' button. The first row has a green checkmark icon.

To the left of the table, there's a sidebar with a 'Search and filter' section and three progress bars: 'Salesforce 99%', 'Markoto 50%', and 'NetSuite 29%'. On the right side, there are several panels: 'OCCUPATION' (with a 'COLUMN' tab), 'TASK MANAGEMENT' (with 'Find a function...', 'Split the task', and 'Assign source's value to golden record' buttons), and 'CHART' and 'PATTERN' sections with various data visualization tools.

» Figure 10 : Effectuer des tâches de remédiation des données avec Talend Data Stewardship

Donnez à vos collaborateurs les moyens de protéger et masquer les données

L'équipe de gouvernance des données peut également déléguer des responsabilités en matière de protection des données, par exemple, le masquage des données. Dans un data lake, par exemple, il est possible que les spécialistes IT ne soient pas responsables du masquage des données, voire qu'ils n'aient pas les priviléges d'accès pour traiter les données avant qu'elles ne soient masquées. Une fois encore, vous devrez ensuite pouvoir déléguer la protection des données à des collaborateurs non spécialistes du masquage des données ni experts techniques.

C'est pourquoi il est essentiel de donner l'autonomie nécessaire à un large public pour masquer les données. Ainsi, dès l'identification de scénarios susceptibles de révéler des données sensibles, ces intervenants peuvent intervenir de façon proactive à l'aide d'un outil convivial. C'est pour cela que la plate-forme Talend permet le masquage des données dans l'ensemble de ses applications, de Talend Data Catalog à Talend Data Preparation, en passant par Talend Studio.

Étudions le cas d'usage suivant. Un responsable de campagne prépare un événement avec un partenaire, mais les clients n'ont pas consenti explicitement au partage de leurs données personnelles avec des tiers. Heureusement, Talend Data Preparation propose un ensemble d'actions de masquage des données par simple glisser-déposer. Le responsable de la campagne pourra donc masquer directement ces données et assurer un partage des données simple et conforme aux règles sur la confidentialité des données.

Nous sommes maintenant arrivés au terme de la 2e des 3 étapes de l'approche permettant d'obtenir des données fiables. Les données sont désormais disponibles depuis un point d'accès unique et peuvent être rapprochées afin que vous puissiez comprendre les relations et le lignage entre les ensembles de données. Vous pourrez ensuite définir les responsabilités pour ces ensembles de données et permettre aux propriétaires de compiler des données, de les corriger ou de les protéger, ou déléguer ces tâches à d'autres spécialistes des données ou utilisateurs métiers. Vous venez de créer un point unique de données fiables et toujours exactes pour vos ressources de données.

» Figure 11 : Masquage des données avec Talend Data Preparation

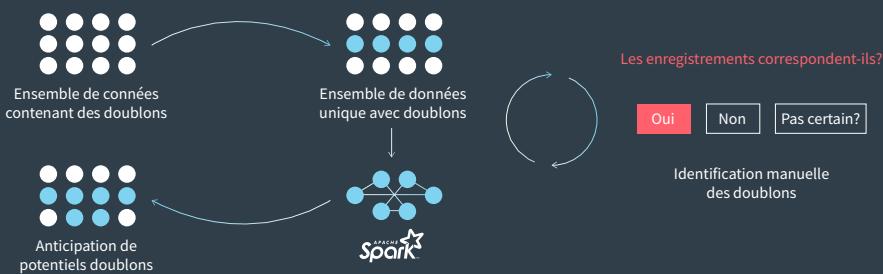
3^e étape : automatisez vos pipelines de données et autorisez l'accès aux données

Vous contrôlez désormais parfaitement vos données. Il est donc temps d'en exploiter pleinement la valeur en permettant à un large panel de personnes et d'appareils autorisés d'y accéder.

L'automatisation est incontournable à l'ère du numérique. Dans la deuxième étape, nous avons vu qu'il était capital de faire participer les collaborateurs au processus de gouvernance des données. Toutefois, ils risquent de constituer des points de ralentissement (rappelez-vous la métaphore de l'Encyclopædia Britannica). Par conséquent, il convient de développer leurs compétences, de les libérer des tâches répétitives et de veiller à ce que les politiques qu'ils ont définies s'appliquent à tous les flux de données, de manière systématique. Des technologies telles que l'intégration des données et le machine learning peuvent être utiles en ce sens.

Exploitez la puissance de l'automatisation pour rationaliser vos flux de données. Utilisez le machine learning pour tirer des enseignements de la remédiation et vous adapter plus rapidement.

Rationaliser les flux de données avec l'automatisation



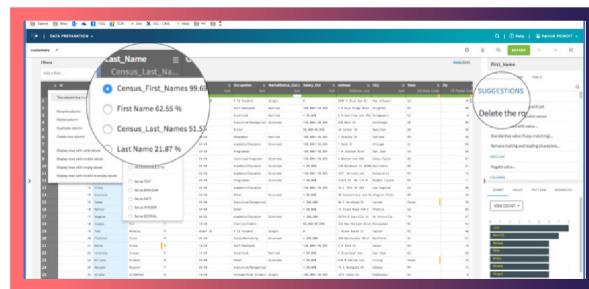
L'analytique avancée et le machine learning simplifient grandement les tâches et contribuent ainsi à démocratiser la gouvernance et la gestion des données. Ils augmentent la productivité des développeurs, suggèrent la meilleure action à effectuer et permettent aux non-spécialistes d'exploiter les données en se laissant simplement guider. Prenons un exemple avec Talend Data Preparation ; lorsqu'en inspectant les données, le logiciel établit qu'il pourrait être utile d'uniformiser leur valeur dans une colonne de texte, il propose à l'utilisateur d'essayer la fonction « find and group similar text » (trouver et regrouper les textes semblables). Cette fonction utilise une fonctionnalité avancée de qualité des données appelée text clustering. Grâce à l'utilisation de fonctions intelligentes telles que la reconnaissance ou des recommandations de structure, le logiciel peut mettre cette fonction à disposition d'utilisateurs non avertis.

Le machine learning permet également de recueillir les connaissances des utilisateurs métiers et des spécialistes des données. La résolution des erreurs et le recouplement des données constituent des exemples types. Avec des outils en libre-service comme Talend Data Stewardship, il est possible de dédupliquer des enregistrements dans un échantillon de données puis d'appliquer le machine learning à l'ensemble du dataset au sein d'un processus entièrement automatisé. Ainsi, Talend transforme des tâches chronophages à faible valeur ajoutée en processus automatisé utilisable à grande échelle sur des millions d'enregistrements.

Le machine learning suggère la meilleure action à effectuer au niveau du pipeline de données ou recueille les connaissances tacites des utilisateurs de la plate-forme Talend (développeurs dans Talend Studio ou stewards dans Talend Data Stewardship) pour automatiser les processus en fonction des besoins.

Notre objectif est de placer les données au centre de toutes nos activités. Toutes les actions courantes réalisées, les engagements auprès des clients et les décisions prises s'appuient sur du big data et de l'analytique.

Adrian Vella, responsable data and business intelligence chez Tipico



» Figure 13 : Assistance intelligente grâce au machine learning dans Talend Data Preparation

Automatissez la protection avec un masquage des données toujours disponible

Le masquage des données vous permettra de partager des données de qualité de manière sélective avec l'ensemble de votre entreprise à des fins de développement, d'analyse, etc., sans divulguer de données personnelles à des personnes n'étant pas autorisées à les consulter.

Si elle est incapable de garantir la protection des données, votre entreprise sera vulnérable, sa réputation sera compromise et vous encourrez des sanctions réglementaires. Pour y remédier, vous devez trouver un moyen pour repérer automatiquement les ensembles de données sensibles et c'est justement l'une des fonctionnalités clés des technologies de catalogage des données.

L'automatisation du processus d'identification des données personnelles débute généralement par la création d'un catalogue de données. Une fois que vous aurez défini les données personnelles parmi vos éléments de données, vous pourrez repérer automatiquement les ensembles de données s'y rapportant. Le masquage de ces éléments peut permettre de réduire les contraintes en matière de conformité réglementaire. Plus la protection sera importante, plus la gestion de vos données sera sécurisée. Si des données personnelles ne présentent

aucun intérêt pour les tests ou l'analytique, pourquoi courir le risque de les révéler ? Grâce au masquage, vous minimiserez les risques tout en protégeant les données. Vous anonymiserez les ensembles de données de sorte que les personnes ne soient plus identifiables.

Par le passé, le masquage des données était réservé à quelques initiés. Alors que les scandales autour de la protection des données et les réglementations se multiplient, vous devez mettre en place un masquage des données plus généralisé que vous pouvez intégrer dans votre flux de données. C'est la condition sine qua non pour pouvoir partager des données de qualité dans l'ensemble de votre entreprise à des fins de business Intelligence, sans divulguer de données personnelles.



» Figure 14 : Ajout d'une protection permanente des données

Point sur le RGPD :

L'article 25 du RGPD définit la protection des données dès la conception et par défaut, tandis que le considérant 26 indique que les principes de protection des données doivent s'appliquer à toute information concernant une personne physique identifiée ou identifiable. Les lois en matière de protection des données ne doivent par conséquent pas s'appliquer aux informations anonymes, c'est-à-dire celles qui ne se rapportent pas à une personne physique identifiée ou identifiable ou à des données à caractère personnelles rendues anonymes de telle manière que la personne concernée ne soit pas ou plus identifiable.

Les responsables de la gouvernance des données doivent établir des contrôles afin de masquer ou de crypter les données personnelles sensibles de façon adéquate. Les normes de masquage des données doivent garantir que les données ne pourront pas être reconstituées lorsque plusieurs champs seront associés. Par exemple, les data scientists peuvent demander que le champ du nom de l'employé soit masqué avant toute analyse. Toutefois, un data scientist perspicace pourra probablement déduire l'identité en associant son poste, sa rémunération et son sexe (par exemple, « responsable des RH, femme, avec un salaire de base de 200 000 \$ »). Dans ce cas, il peut être plus judicieux de masquer l'intitulé du poste et de fournir uniquement une fourchette de salaire comme par exemple « plus de 100 000 \$ ».

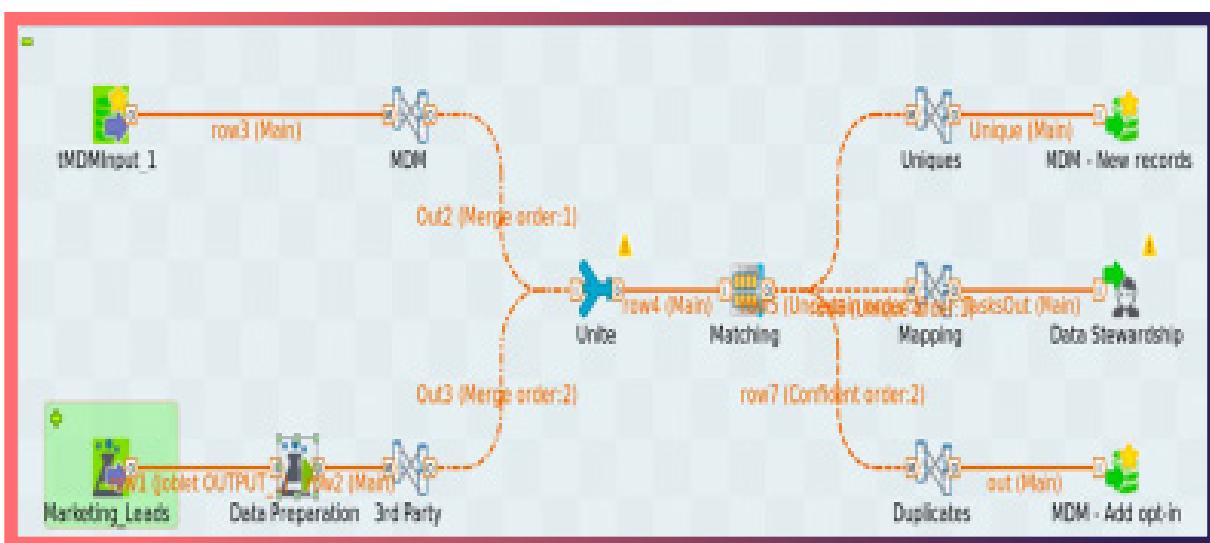
Automatisez vos pipelines de données

Dans de nombreux cas, l'approche de gouvernance des données échoue car elle ne peut pas être appliquée de façon systématique. Prenons l'exemple d'un inventaire des données. Des études montrent que, la plupart du temps, les inventaires des données sont créés selon une approche déclarative basée sur des entretiens avec les propriétaires de données et de processus et sur une documentation utilisant des outils par formulaire ou Excel. La création de

cet inventaire des données est un processus rapide qui exige d'importantes ressources humaines, mais devient obsolète dès que l'environnement des données évolue.

C'est pourquoi des contrôles modernes de gouvernance des données doivent être intégrés à la chaîne de données de façon à ce qu'ils soient opérationnels sans pouvoir être contournés. Ils permettent aux ingénieurs de données de coordonner et d'automatiser tous vos pipelines de données entre n'importe quel entrepôt de données (dans le cloud ou sur site) et n'importe quelle application ou analyse. Ils orchestreront l'opérationnalisation et l'automatisation de tout job ou flux afin que vous puissiez poursuivre la structuration et le nettoyage de vos données tout au long de leur cycle de vie, pendant que des stewards assurent la validation, des utilisateurs la compilation et des utilisateurs métiers la préparation des données.

Chez Talend, Talend Studio est la clé de voûte de tous les flux de données : il offre une large gamme de fonctionnalités techniques couvrant l'intégration, le profilage des données et de nombreux contrôles de qualité des données. Il permet d'opérationnaliser des tâches intégrées dans des outils en libre-service tels que Talend Data Preparation.



» Figure 15 : Automatisation des pipelines de données avec Talend Studio

Générez des données fiables grâce à un accès simple via une recherche à des données de qualité

L'obtention de données fiables permet aux entreprises de recueillir les données de façon centrale, de préserver leur exactitude et de les publier selon des règles et des politiques précises. L'avantage de cette approche est qu'elle contrôle les données, mais aussi qu'elle les libère afin qu'elles puissent être utilisées. Elle permet aux spécialistes des données de trouver, comprendre et partager des données dix fois plus vite. Le temps que les ingénieurs de données, les data scientists, les analystes et même les développeurs ne consacrent pas à rechercher ou recréer les ensembles de données peut servir à exploiter ces ensembles. Dès lors, votre data lake ne risque plus de se transformer en marécage.

Un catalogue de données ne permet pas seulement aux propriétaires des données de les compiler et les gérer. Il les rend également plus pertinentes pour les utilisateurs grâce à ses fonctionnalités de profilage, d'échantillonnage et de catégorisation, de documentation des relations entre les données et de collecte participative des commentaires, des tags, des mentions j'aime et des annotations. Toutes ces métadonnées peuvent ensuite être utilisées très facilement par le biais d'une recherche en texte intégral ou par facettes ou via la visualisation des flux de données. Talend Data Catalog accélère la recherche, l'utilisation et l'accès à des données fiables en vérifiant leur validité avant de les partager. Par son caractère participatif, il transforme les consommateurs de données en data curators et leur permet de fournir des métadonnées ou des informations destinées au glossaire métiers.

The screenshot shows a search interface for 'Payable account'. The search bar at the top contains the text 'Payable account'. Below the search bar, there is a 'Search Filters' section with two main categories: 'Object Types' and 'Models'. Under 'Object Types', 'All' is selected, and under 'Models', 'Enterprise Glossary' is selected. The results list shows several entries, each with a small icon, the term name, and a detailed description. The results are as follows:

- Payable Account (Term) of Enterprise Conceptual Model in Enterprise Glossary (Development)
- Finance Account (Term) of Enterprise Conceptual Model in Enterprise Glossary (Development)
- Account Name (Term) of Enterprise Conceptual Model in Enterprise Glossary (Development)
- Account Description (Term) of Enterprise Conceptual Model in Enterprise Glossary (Development)
- Account Identifier (Term) of Enterprise Conceptual Model in Enterprise Glossary (Development)
- Finance Account Amount (Term) of Enterprise Conceptual Model in Enterprise Glossary (Development)
- General Ledger Account (Term) of Enterprise Conceptual Model in Enterprise Glossary (Development)
- Account (Term) of Enterprise Conceptual Model in Enterprise Glossary (Development)
- Receivable Account (Term) of Enterprise Conceptual Model in Enterprise Glossary (Development)

At the bottom right of the results list is a blue 'Filter' button.

» Figure 16 : Accès aux données via une recherche

À retenir :

Dans son rapport « Data Intelligence Software for Data Governance », IDC vante les avantages d'une gouvernance moderne des données et considère Data Catalog comme la clé de voûte de ce qu'il définit comme un logiciel de data intelligence. Dans ce rapport, IDC la définit comme une « technologie qui favorise l'implémentation par le biais de la gouvernance et d'un logiciel de data intelligence et précise qu'elle est proposée dans des logiciels de gestion des métadonnées, de lignage des données, de catalogage des données, de création de glossaires métiers, de profilage des données, de mastering et de gestion. ».

Donnez à tous la possibilité d'agir : mettez en place une plate-forme unique pour tous et tirez parti d'applications conviviales destinées à votre communauté d'acteurs

Un catalogue de données aidera grandement les consommateurs de données à trouver leurs données, mais l'expérience ne s'arrête pas là. Maintenant que vous avez trouvé les données, il est temps de les exploiter. Pour cela, les collaborateurs ont besoin d'applications faciles à utiliser et adaptées à leur rôle. Mais les applications universelles qui répondent aux besoins de chacun (analyste, ingénieur de données ou encore développeur informatique) n'existent pas. Le cloud offre un modèle de déploiement idéal pour ces applications prêtes à l'emploi qui renvoient vers n'importe quel ensemble de données, quel que soit son emplacement. Les architectures de données telles que [Talend Cloud](#) proposent un ensemble complet d'applications collaboratives qui facilitent l'utilisation des données fiables obtenues au terme des trois étapes présentées ici.

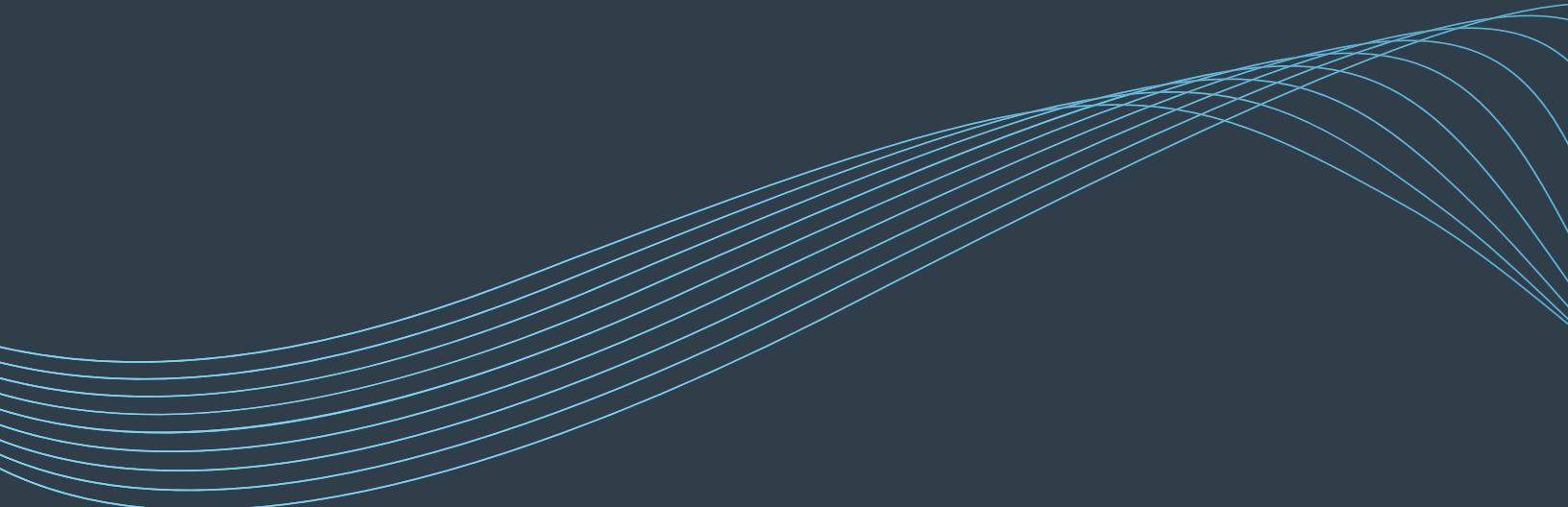
Les services de données publient les données fiables au sein d'applications

Parallèlement à votre projet de gouvernance des données, vous avez maintenant créé des données prêtes à être réutilisées. Elles ne doivent pas être réservées aux utilisateurs métiers de votre entreprise ; les applications métiers devraient également pouvoir en profiter. Ces applications sont très diverses : services financiers de paiement, campagnes marketing coordonnées par Marketo ou applications d'apprentissage qui explorent votre profil pour vous proposer les meilleurs programmes. Identifiez les applications qui nécessitent des données fiables et connectez-les aux très précieux ensembles de données établis dans le cadre de votre programme de gouvernance des données. Tous les pipelines de données créés et automatisés sont non seulement destinés aux tableaux de bord de la business Intelligence, mais aussi aux applications qui en tireront pleinement profit.

Les API présentent l'avantage considérable de mettre facilement les données à la disposition d'un vaste éventail d'applications. En exploitant des environnements tels que [Talend Cloud API Services](#) dans le cadre de votre plate-forme de gouvernance des données, votre investissement permettra non seulement de bénéficier du potentiel de l'analytique, mais aussi d'alimenter les applications de votre entreprise pour garantir les opérations quotidiennes.

The screenshot shows the Talend Management Console interface. On the left, there is a sidebar with various navigation options: OPERATIONS, MANAGEMENT, PROJECTS, ENGINES, ENVIRONMENTS, PROMOTIONS, CONFIGURATIONS, USERS, GROUPS, ROLES (which is highlighted with a green circle), and SUBSCRIPTIONS. The main area is titled 'MANAGEMENT CONSOLE' and shows a table of roles. The table has columns for NAME, APPLICATIONS, NUMBER OF USERS, and LAST UPDATED. The rows listed are: API Designer (4 users, about 1 month ago), API Tester (4 users, about 1 month ago), Campaign Owner (4 users, 7 months ago), Data Preparation Administrator (3 users, 7 months ago), Data Preparation Manager (1 user, 7 months ago), Data Preparator (2 users, 7 months ago), Data Steward (5 users, 7 months ago), Environment Administrator (3 users, about 1 month ago), Infrastructure Administrator (3 users, about 1 month ago), and Integration Developer (3 users, 7 months ago). At the top of the main area, there is a blue button labeled '+ ADD ROLE' with a yellow circle around it. Above the table, there is a dropdown menu set to 'NAME: 10' and some pagination controls (1/2).

» Figure 17 : Tous les membres de l'entreprise peuvent utiliser des données fiables avec Talend Management Console.



Chapitre 4 :

Recommandations : les 12 travaux de la gouvernance des données

À faire

Définissez clairement vos attentes dès le départ

Oublier ou ignorer la raison d'être des données serait une grave erreur. Ne vous contentez donc pas de gouverner pour gouverner. Qu'il s'agisse de minimiser les risques ou d'optimiser vos bénéfices, associez votre projet de gouvernance des données à des résultats clairs et mesurables. La gouvernance des données étant une initiative mise en place à l'échelle de l'entreprise et pas simplement des départements, vous devrez prouver son intérêt dès le départ afin de convaincre les responsables de la prioriser et de lui attribuer des ressources.

Cherchez-vous à rejoindre Ithaque ? Réfléchissez à votre définition de la réussite

Pour Ulysse, Ithaque est la destination finale, la fin de son Odyssée. La réussite peut revêtir différentes formes : renforcer le contrôle des données, limiter les risques ou les violations de données, réduire le temps passé par les équipes métiers, monétiser les données ou produire une nouvelle valeur à partir de vos pipelines de données. Il est essentiel de réfléchir au respect des normes de conformité pour éviter des sanctions.

Trouvez votre financement

À mesure que vous élaborerez les principes fondamentaux de vos projets et que vous définirez les critères de réussite, vous devrez expliquer le « quoi, pourquoi et comment » de la gouvernance des données. N'oubliez pas non plus le « combien ? ». Identifiez les coûts connexes et les ressources impliquées. Si vous occupez depuis peu le poste de délégué à la protection des données, assurez-vous que vous disposez déjà d'un fond de fonctionnement minimum. Si vous êtes responsable des données ou CDO, alliez-vous à la direction informatique

pour obtenir ensemble vos financements. Défendez ensuite votre proposition auprès de votre équipe financière afin qu'elle comprenne les risques de non-conformité auxquels s'expose l'entreprise, expliquez l'intérêt de votre stratégie de données et le potentiel caché des données. Assurez-vous de lui faire prendre conscience que les données sont un actif financier.

Entourez-vous

Comme vous le savez – et nous ne le dirons jamais assez –, une transition vers la gouvernance des données n'est pas juste un projet comme tant d'autres devant être porté par le département IT.

Même si vous pouvez rapidement maîtriser les outils et tirer parti d'applications puissantes, la restitution de données fiables est un travail d'équipe. Réunissez vos collègues des différents départements et mettez en place un groupe de discussion sur les enjeux que les données présentent pour eux. Essayez d'identifier le type de problèmes auxquels ils sont confrontés. Les griefs fréquemment soulevés sont les suivants :

- « J'ai du mal à accéder aux ensembles de données. »
- « Je n'arrive pas à trouver les données que je recherche. »
- « Les données de Salesforce sont polluées. »
- « Comment puis-je m'assurer qu'elles sont fiables ? »
- « Nous passons trop de temps à supprimer manuellement les doublons. »

Vous prendrez bientôt conscience que l'un des plus grands défis consiste à bâtir une chaîne de valeur des données dont des profils très variés peuvent tirer parti pour obtenir des données plus faibles dans les pipelines de données. Travaillez avec vos homologues pour définir les problèmes, les documenter et voir ensemble comment vous pouvez les éliminer. Intéressez vos collaborateurs à votre transition vers la gouvernance des données et donnez-leur des responsabilités afin qu'eux aussi s'approprient le projet. Démontrez-leur que la réussite du projet bénéficiera à l'ensemble des membres de l'équipe.

Intéressez vos collaborateurs à votre transition vers la gouvernance des données et donnez-leur des responsabilités afin que votre projet devienne aussi leur projet.

Positivez votre gouvernance

Dans la mesure du possible, évitez d'exercer un contrôle excessif et une approche descendante trop autoritaire. En revanche, appliquez le modèle collaboratif et maîtrisé de la gouvernance des données dès le départ. Ainsi, vous pourrez mettre en place des applications contrôlées prenant en compte les rôles de chacun pour que vos acteurs et l'ensemble de leur communauté puissent exploiter la puissance des données.

Assurez-vous que l'entreprise comprend les avantages d'une telle gouvernance et qu'elle est prête à collaborer pour fournir des données fiables en temps réel.

Commencez par vos données

La gouvernance traditionnelle prévoit souvent d'appliquer une approche descendante non négociable à l'attribution des responsabilités concernant les données. Vous devez consacrer du temps à déterminer les axes de votre gouvernance des données. Il est évident que cette phase ne sera pas très productive, car vous vous heurterez souvent à une grande résistance. Commencez plutôt par vos données et ceux qui les utilisent. Les entreprises orientées données commencent à se lancer dans la découverte de leurs données. Les spécialistes des données doivent écouter les experts métiers et leurs collaborateurs, explorer les ensembles de données pour repérer leur valeur commerciale et les éventuels risques financiers, puis identifier ceux qui utilisent le plus ces ensembles de données. Les utilisateurs avancés seront souvent les plus enclins à protéger vos ensembles de données, à les corriger et à garantir un haut niveau d'intégrité.

La gouvernance des données n'est pas un simple projet, c'est un processus continu.

Nitin Kudikala, Customer Success Architect chez Talend

Optez pour des solutions cloud

Gartner [prévoit](#) qu'« à l'horizon 2023, 75 % de l'ensemble des bases de données seront dans une plate-forme cloud, ce qui complexifiera la gouvernance et l'intégration des données ». La transition vers le cloud s'accélère : les entreprises doivent collecter des volumes de données croissants, y compris de nouveaux ensembles de données créés derrière leurs pare-feux et mettre ces données à la disposition de publics plus vastes en temps réel. Elles aspirent à plus d'agilité et à des fonctionnalités de traitement à la demande.

Vos données pouvant se trouver hors site dans des infrastructures tierces, le cloud pourrait exiger la mise en place des principes de gouvernance des données plus forts. Prenons l'exemple de la protection des données.

La réglementation en la matière impose que :

- vous mettiez en place des contrôles pour l'échange transfrontalier de données ;
- vous élaboriez des politiques de notification des violations de données et que vous mettiez en place des principes de protection clés (portabilité des données, politiques de conservation ou droit à l'oubli) ;
- vous établissiez des pratiques plus strictes en matière de gestion des relations avec les fournisseurs qui traitent vos données personnelles.

Le cloud est source de nouveaux défis pour vos pratiques de gouvernance des données, mais il ouvre également de nombreux débouchés. Comme nous le verrons dans les cas d'usage, les clients de Talend choisissent, pour l'essentiel, pour le cloud comme source unique de données fiables. En fonction de votre situation, il est fort probable que le cloud soit la solution idéale ; d'abord pour recueillir toutes les empreintes numériques dans votre environnement de données, et ensuite pour permettre à tous les acteurs de votre processus orienté données ayant recours à des applications prêtes à l'emploi de prendre les rênes et d'utiliser les données.

Préparez-vous à expliquer ce que sont les « données » : ne partez pas du principe que vos interlocuteurs auront la même expertise que vous

Bien souvent, les employés n'ont pas la culture des données. C'est une partie du problème. À mesure que les données prendront de l'importance dans les entreprises, l'ensemble du personnel devra faire l'apprentissage de la culture des données (c'est la « datalphabétisation »). Ils seront également peu enclins à découvrir des outils élaborés. L'utilisation d'un catalogue de données permettra de rendre vos données plus pertinentes, connectées à leur contexte métier et faciles à trouver. Appuyez-vous sur des applications cloud telles que Talend [Data Preparation](#) ou [Data Stewardship](#) pour permettre à vos collaborateurs d'accéder aux données en quelques clics, sans aucune formation particulière préalable.

Prouvez la valeur des données : commencez par des petits projets pour offrir des résultats remarquables

Votre projet se heurtera très probablement au scepticisme. Les sceptiques mettront en doute votre capacité à maîtriser et à résoudre leurs problèmes. Et ne partez pas du principe que vos collaborateurs comprendront que vos données ont un intérêt. Vous devrez leur prouver qu'ils économiseront du temps et des ressources en fournissant des données fiables. Prenons l'exemple d'un ensemble de données Salesforce ou d'une source de données Marketo. Utilisez des outils de préparation des données pour expliquer combien il est facile d'éliminer les doublons et d'identifier les problèmes de qualité des données. Présentez la fonction de « recettes » qui permet de répéter le travail de préparation pour d'autres ensembles de données. La qualité des données vient en premier. Mais, plus encore, assurez-vous qu'ils comprennent les avantages qu'elle présente puisqu'ils s'appuient sur des coordonnées de clients dont la qualité a été contrôlée pour améliorer le retour sur investissement de leurs activités de vente et de marketing.

Pour les convaincre rapidement, vous pouvez également leur démontrer combien il est facile de masquer les données avec Talend Data Preparation.

Talend Cloud Data Preparation possède une interface très simple et facile à utiliser qui nous permet de valoriser nos données beaucoup plus rapidement.

Jermaine Ransom, vice-président des services de données chez DMD Marketing Corp

À ne pas faire

Sachez que le soutien de l'exécutif ne sera aucunement garanti

Une fois que vous aurez prouvé l'intérêt économique par de petites démonstrations du concept et obtenu le soutien de l'entreprise, demandez à rencontrer votre direction. Présentez-lui votre projet pour améliorer les données dans l'ensemble de l'entreprise. Soyez clair, concis et bref afin que chacun puisse comprendre l'intérêt de votre projet. Expliquez-leur qu'ils gagneront en visibilité en vous soutenant et, par conséquent, qu'ils amélioreront l'efficacité de l'entreprise dans son ensemble.

Vous gagnerez leur confiance qui est indispensable pour qu'ils soutiennent votre projet, et votre travail n'en sera que simplifié.

Adoptez une approche pratique et non passive. Dirigez votre projet de données fiables

Rencontrez différents collaborateurs pour en savoir plus sur les défis auxquels ils sont confrontés et leur offrir votre aide ; ils vous considéreront comme le chef du projet. Assurez-vous que toutes vos actions sont efficaces. Planifiez le travail et tenez-vous-en au plan. Suivez chaque étape du projet et présentez les étapes suivantes. Vous serez confronté à des obstacles et devrez remanier vos priorités à mesure que votre entreprise s'adaptera aux évolutions des conditions du marché. N'abandonnez pas et adaptez votre plan, au besoin. Toutefois, poursuivez vos efforts de persuasion et (ré—)expliquez en quoi votre projet résoudrait les difficultés de l'entreprise.

Assurez-vous que votre gouvernance des données est bien en lien avec vos données. Un trop grand nombre de programmes de gouvernance des données ont mis en place des politiques, des workflows et des procédures sans les associer aux données. Par exemple, [une enquête a montré](#) que, sur les 98 % d'entreprises interrogées qui affirmaient respecter le RGPD dans leurs mentions légales, seuls 30 % étaient en mesure de répondre effectivement aux demandes d'accès aux données des clients exerçant leurs droits en la matière. Cela montre bien que la plupart des entreprises ont fixé des principes de gouvernance forts, mais ne les mettent pas en application.

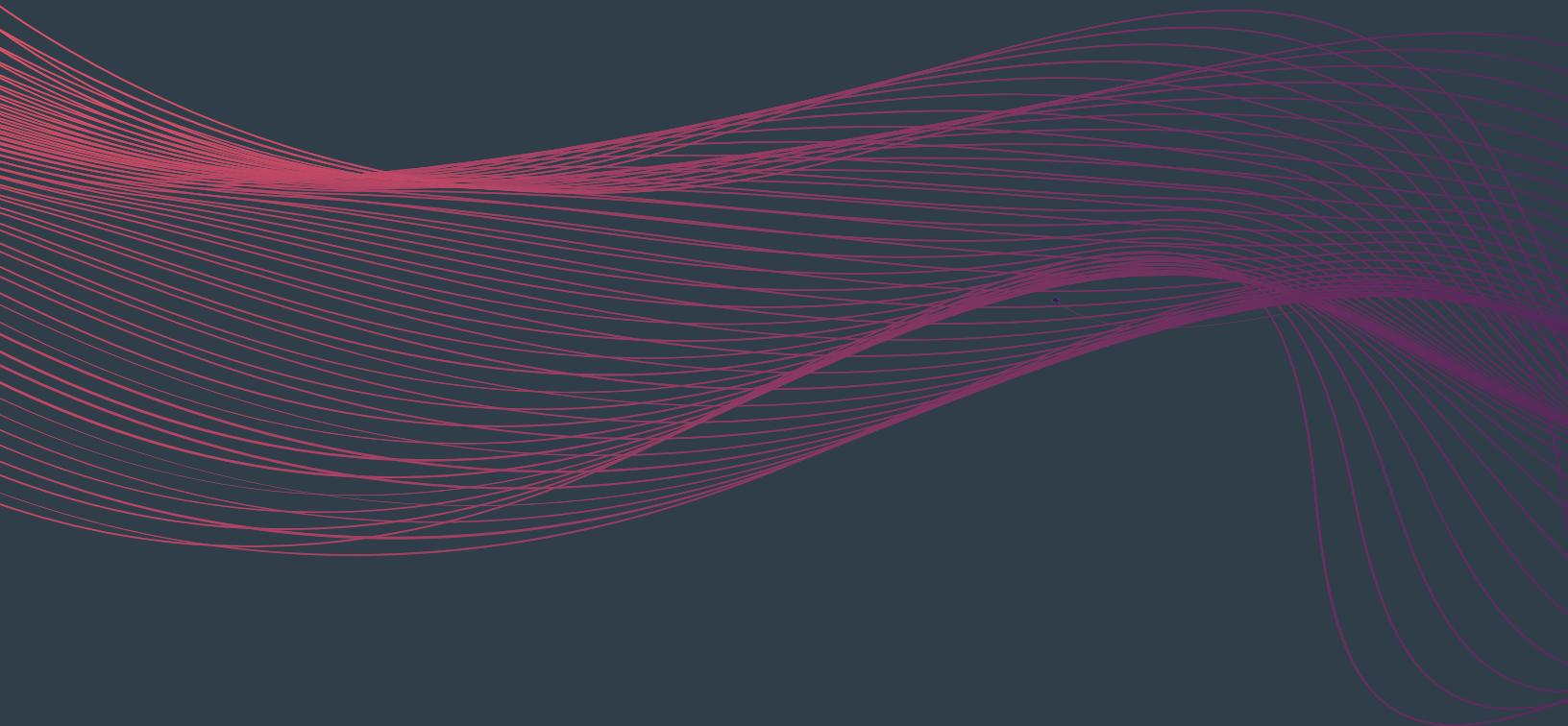
Relevez le défi de vos données.

Imaginons une crise en situation réelle dans laquelle vous avez subi une violation ou une fuite de données en interne. Vérifiez si votre cadre de gouvernance des données est efficace dans le cas d'un scénario catastrophe. Réalisez une piste d'audit. Toutes vos données sensibles sont-elles masquées ? Êtes-vous en mesure de suivre toutes vos données et d'en garantir la traçabilité ? Les propriétaires de données se sentent-ils responsables des données dont ils ont la charge ? Mettez-vous à la place de votre client. Pensez à son droit d'accès aux données ou à son droit à l'oubli.

Pourquoi ne pas réaliser une simulation en équipe ? Préparez un scénario et vérifiez dans quelle mesure votre plan est efficace, puis appuyez-vous sur les enseignements de cette simulation pour l'améliorer. Mieux vaut prévenir que guérir. Donc, soyez proactif pour ne pas vous retrouver à gérer une violation réelle de la confidentialité des données, avec toutes ses conséquences. Ainsi, votre gouvernance des données sera plus concrète, basée sur des défis d'ordre opérationnel plutôt que sur de grands principes.

Chapitre 5 :

Les nouveaux rôles de la gouvernance des données



Parlons maintenant des rôles que vous devriez avoir au sein de votre équipe de gestion des données et de leur place dans le cadre de l'approche collective de la gestion des données.

Gardez à l'esprit qu'il n'y a pas de modèle universel parfait d'équipe de gestion des données. En fonction de la culture de votre entreprise et de la façon dont elle perçoit les risques, cette équipe peut être davantage orientée sur le risque que sur la valeur. Cela dit, votre équipe doit disposer des compétences et des savoir-faire spécifiques pour pouvoir comprendre les réglementations en matière de conformité et de gestion des données. Ainsi vous pourrez couvrir toute la palette des données nécessaires à la mise en œuvre de votre stratégie en la matière.

Le fait est qu'au cours des dernières années, les rôles liés aux données ont connu une transformation radicale dans la plupart des entreprises. Au sein du département Applications métiers, ces postes autrefois centralisés deviennent clairement de plus en plus décentralisés, un signe évident que les entreprises commencent à adopter la collaboration comme moyen de mieux gérer les données. Gartner souligne que « les rôles clés, tels que le data steward, auparavant rattachés au groupe IT évoluent vers une affectation unique aux unités métiers ou une combinaison hybride IT-métiers. »

En règle générale, un cadre de gestion collaborative des données comporte six rôles essentiels :

Les directeurs des données ou CDO (Chief Data Officer) : ils sont les chefs d'orchestre de la stratégie de données. Ils sont chargés de définir, de déployer et de suivre la stratégie de données avec l'aide de l'équipe de gouvernance des données. Leur mission consiste à s'assurer que les données sont considérées comme des ressources métiers de valeur

au niveau exécutif. Grâce à cela, le comité de direction investit dans des ressources pour garantir la conformité des données en vue de minimiser les risques et de valoriser les flux de données pour optimiser les revenus. Certains deviennent directeur de la confiance numérique (Chief Trust Officer) après avoir transformé des données en ressources fiables avec des résultats positifs pour l'entreprise.

Les responsables de la protection des données ou DPO (Data Protection Officers) : le rôle du responsable de la protection des données est crucial, car sa mission consiste à assurer le respect des normes en matière de protection des données définies par les autorités compétentes. Les responsables de la protection des données s'assurent que les données personnelles traitées par l'entreprise sont parfaitement conformes. Exemples : données sur les clients, les fournisseurs ou toute autre personne, traitées par l'entreprise dans le cadre de ses activités quotidiennes. Leur rôle consiste à s'assurer que les données sont protégées conformément à certaines lois et réglementations de la région géographique où leur entreprise opère. Ils doivent être en relation directe avec le comité de direction conformément à certaines réglementations, notamment le RGPD ou la CCPA.

Le fait est qu'au cours des dernières années, les rôles liés aux données ont connu une transformation radicale dans la plupart des entreprises.

Les data architects : leur rôle est vital. Comme le dit l'adage : « C'est au pied du mur qu'on voit le maçon. » La logique est la même pour les architectes de données. Ils sont chargés de s'assurer que l'entrepôt de données est solide et robuste, tout en étant suffisamment flexible pour que les utilisateurs puissent devenir les maçons de leur propre dataset lorsqu'ils annotent, enrichissent ou certifient des données. Ils définissent les fondements des données-ressources afin qu'elles soient pertinentes et orientées métier pour l'ensemble de l'entreprise. Ils peuvent hiérarchiser les informations en fonction de résultats mesurables.

Les data stewards : ils sont chargés de garantir l'intégrité des données dans des ensembles de données précis. Par exemple, le directeur du CRM peut être chargé de la gestion de la base de données de clients. La responsabilité des data stewards est grande ; ils doivent s'assurer que leurs ensembles de données respectent des normes de qualité définies par l'équipe de gouvernance des données. Ils peuvent travailler en partenariat avec d'autres « information stewards » pour déployer le processus d'intégrité des données dans leur secteur géographique/division/département.

Les ingénieurs de données et développeurs : les ingénieurs de données sont chargés de concevoir, de déployer et de maintenir une architecture qui permet le traitement de flux de big data et de données complexes au sein de l'entreprise. Ils ont une formation technique et utiliseront des environnements de données techniques pour traiter ces flux de données. Initialement, ils étaient

chargés du mouvement des données, mais ils doivent désormais connaître le contenu du pipeline et s'assurer qu'il a fait l'objet de contrôles de qualité. Souvent submergés par un nombre croissant de demandes provenant des utilisateurs métiers, de plus en plus d'ingénieurs de données s'efforcent de conférer plus d'autonomie à l'ensemble de la communauté d'utilisateurs de données, tout en gardant le contrôle des accès, des autorisations et de l'administration des données.

Les data scientists : les data scientists extraient la valeur des pipelines de données afin de fournir de précieuses informations à l'entreprise. Les data scientists résolvent des problèmes compliqués liés aux données en s'appuyant sur les mathématiques, la statistique et l'informatique. Ils sont souvent experts en statistique, en data mining et en analyse prédictive. Leur expertise s'accompagne parfois de compétences en programmation.

Les business analysts : leur mission consiste à étudier les tendances et les opportunités, identifier et calculer les risques et prendre le pouls de l'entreprise. Ils utilisent intensivement les outils décisionnels tels que Tableau, Power BI ou Qlik. Ils doivent extraire des informations fiables des pipelines de données afin de les présenter sous forme résumée dans des tableaux de bord complets.

À retenir :

Plus vous ciblez des professionnels, plus les applications en libre-service doivent être simples et intelligentes.

Utilisateurs métiers/data curators : venant de tous les départements, ils ne sont ni experts ni spécialistes. Leurs compétences et leurs capacités peuvent différer, mais ils déplorent la difficulté d'accès aux données tout en étant soucieux d'exploiter la valeur des données à l'aide d'outils en libre-service simples et modernes.

Il est essentiel qu'ils puissent travailler en équipe. La gouvernance des données collaborative est un sport d'équipe, comme la Coupe de l'America où tous les membres de l'équipage du bateau œuvrent de concert pour gagner la course, unissant leurs compétences et capacités individuelles pour garder une longueur d'avance sur les autres équipes.

Voici un exemple de la manière dont ces rôles collaborent dans la pratique.

Imaginons une entreprise qui souhaite intégrer des données météorologiques pour affiner ses prévisions de vente.

Dans un premier temps, un data scientist saisit les données météorologiques dans un laboratoire de données afin d'affiner le modèle prévisionnel.

Une fois qu'il confirme que ces données influent sur son modèle, un data curator vérifie la qualité, la conformité et les droits d'auteur de ce nouveau dataset.

Le DevOps automatise ensuite l'intégration des données pour ingérer un flux de données en temps réel au sein d'un data lake d'entreprise.

Enfin, les analystes accèdent aux ensembles de données. Ils les partagent avec les utilisateurs métiers, qui attendent ces données, et les interprètent pour eux.

Pour ce projet de données, il est possible d'utiliser une approche cloisonnée dans laquelle le département IT contrôle l'accès aux données et leur distribution à l'aide d'outils d'intégration des données. Les data scientists dans leur laboratoire de données peuvent ensuite utiliser une plate-forme de data science, tandis que les responsables de la gouvernance s'appuient sur ce cadre de gouvernance des données pour garantir la conformité. Toutefois, comment travailler en équipe avec cette approche cloisonnée ? En outre, qui contrôlera cet ensemble disparate de pratiques, d'outils et d'ensembles de données ?

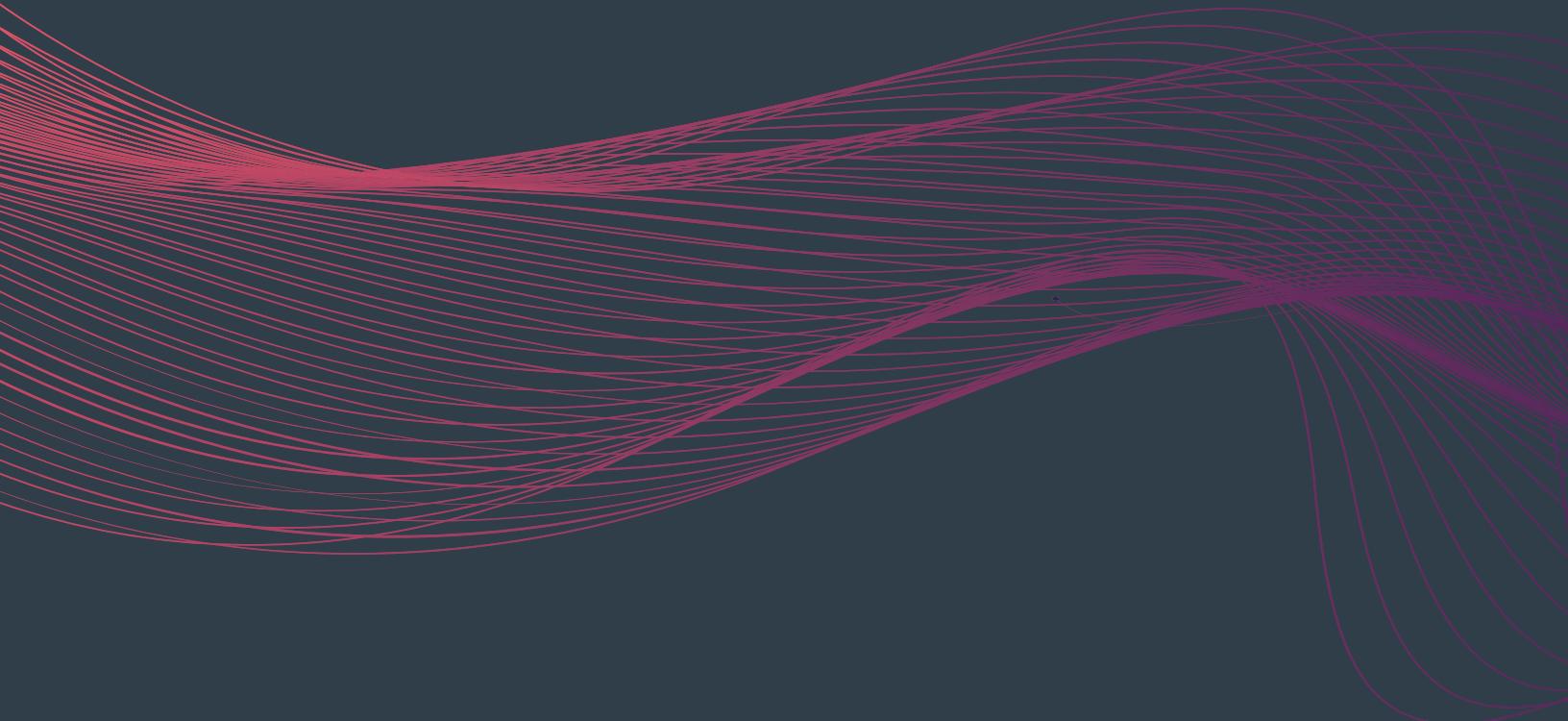
Ce scénario est un exemple parfait de ce qu'est la gestion collaborative des données : permettre un travail en équipe pour exploiter tous les avantages que présentent vos données.

RGPD et protection des données

L'article 25 du RGPD définit la protection des données dès la conception et par défaut, tandis que le considérant 26 indique que les principes de protection des données doivent s'appliquer à toute information concernant une personne physique identifiée ou identifiable. Les lois en matière de protection des données ne doivent par conséquent pas s'appliquer aux informations anonymes, c'est-à-dire celles qui ne se rapportent pas à une personne physique identifiée ou identifiable ou à des données personnelles rendues anonymes de telle manière que la personne concernée n'est pas ou plus identifiable.

Chapitre 6 :

Exemple de réussite en gouvernance de données





Dans le monde du trading, nous avons trois mots d'ordre : l'intégrité, car il est impossible de perdre un seul ordre, la disponibilité permanente et la gouvernance sur un marché très régulé. Talend a répondu à ces attentes.

Abderrahmane Belarfaoui, Chief Data Officer chez Euronext



Euronext devient un « data trader » avec un environnement de données fiables

Après s'être séparée du New York Stock Exchange en 2014, Euronext est devenue la première bourse paneuropéenne de la zone euro, en réunissant les places boursières d'Amsterdam, de Bruxelles, de Dublin, de Lisbonne et de Paris. Euronext comprend près de 1 300 émetteurs représentant une capitalisation boursière totale de 3 700 milliards d'euros à la fin mars 2018.

En 2016, Euronext commence un projet qui a tout d'une classique migration vers le cloud. Sauf que cette migration n'a rien de classique. D'une part, la taille de la base de données que l'opérateur de marchés envisage de mettre dans le cloud est de 100 To. C'est une des plus grandes d'Europe. D'autre part, il ne s'agit pas de simplement la porter sur une plate-forme hébergée. L'idée est de faire un data lake gouverné et de le rendre accessible, en libre-service, aux métiers et aux clients pour monétiser de nouveaux services et générer des revenus additionnels.

Migration vers un cloud gouverné

Avant le lancement de ce projet, Euronext stockait ses données dans un entrepôt de données sur site associé au matériel propriétaire de l'un des grands noms du secteur.

« Notre infrastructure informatique était arrivée en fin de vie dans un contexte européen où les régulateurs souhaitaient qu'Euronext stocke de plus en plus de données », se rappelle Abderrahmane Belarfaoui, Chief Data Officer (CDO) chez Euronext.

« De plus, nous devions parfois attendre six heures – et parfois même davantage – après la clôture des marchés pour pouvoir envoyer les données aux métiers et aux clients. »

Cette situation a amené le CDO à envisager le passage à un modèle cloud hybride permettant de garder une indépendance vis-à-vis du fournisseur du cloud.

Dans un monde extrêmement réglementé, Talend s'est également avéré capable de relever les défis de la gouvernance des data lakes et de la conformité réglementaire. En effet, pour pouvoir ouvrir en toute sécurité les données à de nouveaux usages, notamment la monétisation, et prendre en compte la question de la protection des données (RGPD, etc.), il faut en avoir une connaissance parfaite, avoir la traçabilité des modifications, l'historique des flux de données et savoir comment les classer au sein d'une structure granulaire.

« Nous disposons d'un stockage Amazon S3 qui est partagé avec tout le monde. Je dois savoir qui possède les données (le propriétaire des données), qui a accès à quoi, à qui poser des questions, qui peut les utiliser et qui est prioritaire sur qui. Nos data stewards protègent l'organisation de nos données », ajoute M. Belarfaoui.

Cette stratégie de gouvernance est appliquée dans des outils bien spécifiques, par exemple Talend Data Catalog. Le déploiement d'un dictionnaire se fait en parallèle de chaque projet technique, marché par marché. Ces dictionnaires permettent de retrouver l'historique des données de bout en bout, depuis les sources jusqu'au reporting. « Aujourd'hui, je suis capable de voir qu'une donnée sort de S3, que je lui ajoute une valeur, que je l'agrège avec une autre et qu'elle devient une autre donnée

dans Redshift », se réjouit le CDO, très satisfait du nouveau processus. « Je peux également ajouter des balises. En général, nous ajoutons la durée de stockage. Par exemple, si des données doivent être conservées pendant dix ou cinq ans (conformément à la directive MiFID II) ou si elles doivent être archivées. »

Dans le même temps, le lignage des données avec Talend a considérablement réduit le coût des analyses d'impact. « Un exemple très simple me vient à l'esprit : nous prévoyons de changer la valeur d'un indice de la Bourse britannique. Une fois que nous l'intégrons dans nos systèmes, elle se propage pratiquement partout. Actuellement, il nous faut 200 jours-personnes rien que pour retrouver l'indice dans nos différents systèmes. Mais avec le dictionnaire, nous pouvons lancer le lignage des données d'un seul clic. »

Monétisation des données boursières

Deux ans après son lancement, le projet de data lake gouverné avec Talend et AWS est une réussite. « Les premiers résultats ont été plus que positifs », déclare M. Belarfaoui. « Sur un plan technique, nous pouvons gérer dix fois plus de données d'iso-budgets. »

Au-delà de l'architecture restructurée et de l'amélioration métier dans le cadre de la conformité réglementaire, la nouvelle plate-forme prépare également Euronext à devenir un « data trader ». L'opérateur boursier souhaitait pouvoir affiner la multitude de données dont il dispose et pouvoir la compléter afin de la monétiser. Dans les faits, la monétisation des données représente déjà 20 % du chiffre d'affaires d'Euronext.

Ce projet implique également de donner aux data scientists et aux métiers un accès en libre-service à ces données. Ils pourront ensuite les analyser dans des sandbox de données pour des cas d'usage tels que la surveillance des marchés.

C'est un véritable tournant pour Euronext. « En 2016, nous avions identifié le besoin sans avoir la capacité d'y répondre. À l'époque, nous pouvions uniquement transmettre les volumes de l'activité sur les marchés à des régulateurs de marché (MIFID II). Aujourd'hui, nous pouvons aller plus loin. En vertu du Règlement général sur la protection des données (RGPD), je dois savoir où les données personnelles sont stockées. Si je reçois des demandes de modification ou de suppression, le dictionnaire me permet de trouver les données en question », explique M. Belarfaoui. « De la même façon, un utilisateur qui recherche une transaction peut savoir tout de suite si elle est confidentielle. Une fois que les données sont identifiées comme critiques, le data steward peut en interdire l'accès à certains utilisateurs. »

La gouvernance moderne des données permet aux entreprises de moderniser leur environnement de données

- En garantissant la portabilité des données, accélérer la gestion des modifications et les cycles de migration
- En associant les différents services et le département IT à la création de données fiables et pertinentes
- En retracant les mouvements de données et les activités de traitement dans des systèmes d'information disparates
- En protégeant les données sensibles et en mettant en œuvre des politiques en matière de données
- En offrant à un public plus large des données fiables, de façon contrôlée

Talend est omniprésent dans notre projet. Nous bénéficions des concepts les plus avancés : catalogage des données, informatique sans serveur ou intégration continue.

Abderrahmane Belarfaoui, Chief Data Officer chez Euronext

**SECTEUR D'ACTIVITÉ**

Énergie

INFORMATION :

Siège social : Allemagne

Plus de 10 000 salariés

CAS D'USAGE

Efficacité opérationnelle

PROBLÉMATIQUE

Fournir des données et des analyses en temps réel et en libre-service

PRODUITS TALEND UTILISÉSTalend Real-Time Big Data
Talend Data Catalog
Talend Data Preparation**BÉNÉFICES**

- Intégration de plus de 120 sources internes et externes
- Réduction des coûts d'intégration de plus de 80 %
- Livraison des données en temps réel, dix fois plus vite et dix fois moins cher

ÉCOSYSTÈME DE PARTENAIRES

Azure, Snowflake

**Uniper fournit des données fiables en temps réel**

Uniper produit, négocie et commercialise l'énergie à grande échelle. Avec une capacité de production installée d'environ 36 gigawatts, Uniper compte parmi les plus importants producteurs d'électricité du monde. Uniper achète, stocke, transporte et distribue des produits de base tels que le gaz naturel, le gaz naturel liquéfié (GNL), le charbon et de nombreux autres produits liés à l'énergie.

Parmi les clients Uniper figurent de grands clients industriels et des villes en Allemagne et dans les pays voisins.

Fournir des données et de l'analytique en libre-service et en temps réel

L'industrie de la production et de la distribution d'énergie est au cœur de la plus grande disruption des dernières décennies. La libéralisation de l'énergie entraîne une intensification de la concurrence. Les énergies renouvelables et les technologies de réseau intelligent (smart grid) ont modifié les hypothèses relatives à la planification des investissements, au rapport production centralisée/production décentralisée et aux pratiques de base des spécialistes du secteur.

« Nous évoluons dans un monde de plus en plus complexe dont les marchés et les technologies ne cessent d'évoluer », a déclaré René Greiner, responsable de l'intégration des données chez Uniper SE.

Nous produisons de l'énergie. Nous achetons et vendons de l'énergie sur les marchés. Quelle quantité de charbon et de gaz devons-nous produire aujourd'hui ? Et à l'avenir ? Le marché va-t-il prendre une tout autre direction ? Comment allons-nous développer nos positions sur le marché ? Comment pouvons-nous optimiser nos profits et réduire nos pertes ? Avant d'entreprendre notre transition vers le cloud, nous ne pouvions pas accéder facilement à nos données pour prendre ces décisions rapidement.

« Après avoir formulé l'idée d'une stratégie de données à l'échelle de l'entreprise, nous avons décidé d'opter pour une solution cloud public pour des raisons de coût et de capacité d'évolution. Nous sommes également parvenus à la conclusion que les solutions Talend seraient les mieux adaptées pour une architecture cloud de ce type », explique M. Greiner. « La capacité des solutions Talend à se connecter à un grand nombre de sources et la conception modulaire de leurs produits ont également été des facteurs décisifs. »

Pour prendre des décisions avisées, Uniper s'appuie sur des analyses marketing alimentées par des informations disponibles en temps réel. Maintenant que les informations pertinentes sont agrégées dans le data lake, les équipes d'analyse de marché peuvent accéder plus rapidement aux données et formuler des réponses beaucoup plus rapides aux questions qu'elles reçoivent chaque jour des négociants. Les questions qui exigeaient autrefois plusieurs mois de recherche peuvent désormais recevoir une réponse immédiate ou en quelques jours seulement.

La rapidité de réponse aux questions est essentielle pour les équipes de négociation ; car plus elles peuvent réagir rapidement, plus elles peuvent prendre position avant leurs concurrents, ce qui peut se traduire par une différence de plusieurs millions d'euros.

Pour le gaz par exemple, il est essentiel de comprendre la demande d'électricité et son évolution en fonction de la température, ce qui a un impact immédiat sur les volumes à livrer. Le data lake assure la disponibilité continue des données, ce qui permet aux équipes de négociation d'automatiser les transactions dès que les prix atteignent un seuil prédéfini. Aujourd'hui, Uniper peut prendre position immédiatement, alors qu'auparavant, il fallait plusieurs jours pour rassembler les données nécessaires.

La gouvernance moderne des données permet aux entreprises de réaliser des data lakes et des analyses à plus large échelle

- En accélérant la commercialisation de la découverte initiale à la publication
- En améliorant l'efficacité de la recherche de données et de leur transformation en information
- En mettant en place un contrôle et une protection des données pour les accès en libre-service
- En collectant en mode participatif (crowdsourcing) les connaissances destinées à la compilation des données, aux recommandations et à la remédiation
- En touchant un public plus vaste avec des données fiables et pertinentes

AIR FRANCE KLM

SECTEUR D'ACTIVITÉ

Transport aérien

INFORMATION :

Siège social : France

Plus de 84 000 salariés

CAS D'USAGE

Une approche à 360 ° pour une relation attentionnée à chaque voyageur : « intimité avec le client ».

PROBLÉMATIQUE

Répondre aux besoins spécifiques des clients en matière de voyage

PRODUITS TALEND UTILISÉS

Talend Data Management

Talend Metadata Management

BÉNÉFICES

- Des dizaines de millions d'expériences uniques
- Un million de données traitées tous les mois
- Un accès 10 fois plus rapide pour trouver l'information client

« Notre objectif est de devenir la compagnie aérienne qui répond le mieux aux besoins de ses clients. »

Air France KLM propose des voyages sur mesure avec la plate-forme talend

Air France-KLM est un leader mondial dans ses trois activités principales : le transport aérien de passagers, le fret et la maintenance aéronautique. Avec 90 millions de clients annuels, 27 millions de membres de FlyingBlue et près de 2,5 millions de visiteurs sur Internet chaque mois, le traitement des données clients est une question centrale pour le groupe Air France-KLM.

Répondre aux besoins spécifiques des clients en matière de voyage

Dans le domaine du transport aérien, il va sans dire que la concurrence est vive. Air France-KLM a eu du mal à se démarquer des compagnies low-cost, notamment par ses prix. Il n'a pas non plus été aisément de mettre ses produits en valeur face à ceux de ses concurrents asiatiques et du Golfe. Et le défi ne réside plus vraiment dans la personnalisation de l'expérience du client, mais dans son hyper-personnalisation.

« Nous entrons dans une ère où nous devons répondre aux besoins de chacun de nos clients », explique Gauthier Le Masne, le Chief Customer Data Officer d'Air France-KLM. Aujourd'hui, « les produits ne suffisent plus. C'est par la qualité de notre relation avec nos clients et les services que nous leur offrons que nous nous démarquons de la concurrence. Pour ce qui est des critères de satisfaction des clients, le produit n'occupe que le dixième rang, derrière le service. Les clients ne demandent pas à leur compagnie aérienne de « les transporter » mais plutôt de répondre à leurs besoins en matière de voyage. »

En quelques années, le volume de données à la disposition des compagnies aériennes a décuplé. Les sites et les applications génèrent également de nombreuses interactions. Par exemple, une vente est réalisée toutes les cinq secondes sur airfrance.com. À cela s'ajoute les échanges avec les 16 millions de fans de l'entreprise sur Facebook et ses trois millions de followers sur Twitter, ainsi que des données provenant de campagnes médiatiques, Air France-KLM étant l'un des rares publicitaires à mener son processus d'achats médias en ligne.

Une plate-forme big data pour centraliser les données clients

Si le groupe a commencé à collecter des données clients il y a plusieurs années de cela par le biais de ses centres d'appels, des réseaux sociaux et de son personnel dans les aéroports, les salons d'aéroport et à bord des avions, elles n'ont pas encore été centralisées. Ainsi, le premier défi a été de combiner toutes ces données clients sur une plate-forme commune à toutes les entreprises d'Air France-KLM. « L'idée était que les données de nos clients soient centralisées sur une plate-forme big data afin qu'elles puissent être redistribuées en contexte et en temps réel à l'ensemble de nos points de service à la clientèle », poursuit Gauthier le Masne. La plate-forme a été mise en place au premier semestre 2016. « Nous nous sommes appuyés sur la plate-forme Hadoop que nous avions déjà mise en place. »

En outre, Air France-KLM peut être amené à recueillir et traiter des données personnelles concernant les passagers qui utilisent les services proposés sur son site Internet, son site mobile et dans ses applications mobiles. Avons-nous le droit d'identifier des clients sans leur consentement explicite ? L'entreprise s'est engagée à respecter les réglementations sur la protection des données de ses passagers, des membres de ses programmes de fidélité, de ses prospects et des visiteurs de son site Internet. L'intégralité du traitement des données personnelles est réalisée avec Talend Data Masking, qui permet l'anonymisation de certaines données sensibles et empêche leur identification dans le but de prévenir tout

accès non autorisé. « Talend Data Catalog nous a permis de mettre en œuvre la gouvernance des données, les data stewards et les propriétaires de données étant chargés de documenter les données et les processus », conclut Damien Trinité, CRM Big Data Project Manager chez Air France-KLM. « Air France-KLM peut identifier les données clients, déterminer leur origine et leur destination et partager ces informations au sein de l'entreprise dix fois plus vite qu'avant. »

Améliorer l'expérience de voyage grâce à une approche client à 360°

« Dès qu'il part de chez lui, le client, en déplacement, a besoin d'assistance », explique Gauthier le Masne. Temps de trajet jusqu'à l'aéroport, signalement des retards ou des annulations de vol, temps d'attente à l'aéroport, suivi des bagages. Dès lors, l'entreprise a identifié les principaux facteurs de stress de ses clients pour pouvoir les anticiper autant que possible et être aussi proactive que possible.

Sur le terrain, les agents des centres d'appels ont été les premiers à profiter de cette solution de gestion des données. À bord des avions, tous les chefs de cabine disposent d'un iPad. « Ils ont ainsi accès à toutes les informations concernant les vols et les clients. Concrètement, si un client a l'habitude de prendre un repas végétarien, l'entreprise prendra l'initiative de lui proposer ce menu si l'agent de réservation a oublié de le faire. »

Côté vente et marketing, un moteur de recommandation a été mis en place et des algorithmes orientés données proposent aux clients des tarifs promotionnels sur leurs prochaines destinations préférées.

La gouvernance moderne des données permet aux entreprises de créer une place de marché de données pour leurs ressources les plus partagées

- En créant un inventaire de toutes les données liées à un domaine particulier
- En rapprochant les données disparates au sein d'une vue à 360° de qualité
- En favorisant la propriété des données à des fins de gestion et de compilation des données
- En protégeant les données sensibles contre une utilisation impropre
- En fournissant aux collaborateurs, aux clients et aux applications des données fiables

*Chaque voyageur est unique.
Avec notre plate-forme Big Data
et Talend, nous offrons des
expériences de voyages
« Made-just-for-me » depuis
la réflexion d'achat jusqu'à l'après-vol.*

Gauthier Le Masne, Chief Customer Data Officer chez Air France-KLM

Un cabinet de services financiers réduit son exposition aux risques grâce à un data lake cloud

Utiliser la gouvernance des données pour le rapprochement, l'agrégation et le reporting des risques

La crise financière mondiale de 2007 a eu de graves répercussions sur l'ensemble de l'économie et notamment les établissements de services financiers. Si elle a mis en lumière la nécessité d'adopter des approches plus holistiques pour regrouper les expositions aux risques dans tous les instruments financiers, les activités et les entités légales, elle a aussi imposé de nouvelles exigences réglementaires et entraîné une augmentation des coûts des infractions aux règles en vigueur.

Pour ce cabinet de services financiers, ce défi a conduit à la création d'un service partagé de gouvernance des données centralisée à l'échelle de l'entreprise. Il vise à améliorer la qualité, le rapprochement et le reporting des données grâce à la mise en place de bonnes pratiques en matière de gestion des données.

La modernisation de l'IT était également au cœur du projet. Le cloud et le big data constituent quant à eux les deux piliers permettant de réunir l'ensemble des données sur une même plate-forme d'analytique et de reporting flexible et évolutive. Un data lake sur Amazon Web Services (AWS) lui a permis de recueillir toutes les données brutes nécessaires à l'agrégation des risques.

En utilisant un catalogue de données associé à une gestion de la qualité des données, l'entreprise a pu assurer la fiabilité et la transparence de ce data lake et le diffuser dans l'ensemble de son processus de gestion des risques.

Ouvrir la voie de la confiance, des données brutes aux données conformes

En s'appuyant sur une approche descendante à partir des données brutes ingérées sur un système de fichiers Amazon S3, l'établissement financier a utilisé la plate-forme Talend Real-Time Big Data pour recueillir et intégrer des centaines d'ensembles de données provenant des sources disparates dans le lac de données cloud.

Grâce à Talend Data Catalog, les données qui contribuent à la détermination des risques peuvent être rapprochées, documentées et suivies. Ainsi, les spécialistes des données peuvent consulter le lignage des données de bout en bout, déterminer l'origine des données et les étapes de traitement appliquées pour l'agrégation des risques. En outre, Talend Data Quality génère des rapports sur la qualité et l'exactitude des données sur les risques et favorise la remédiation des données, s'il y a lieu.

Mettre en œuvre la gouvernance des données pour obtenir des résultats mesurables

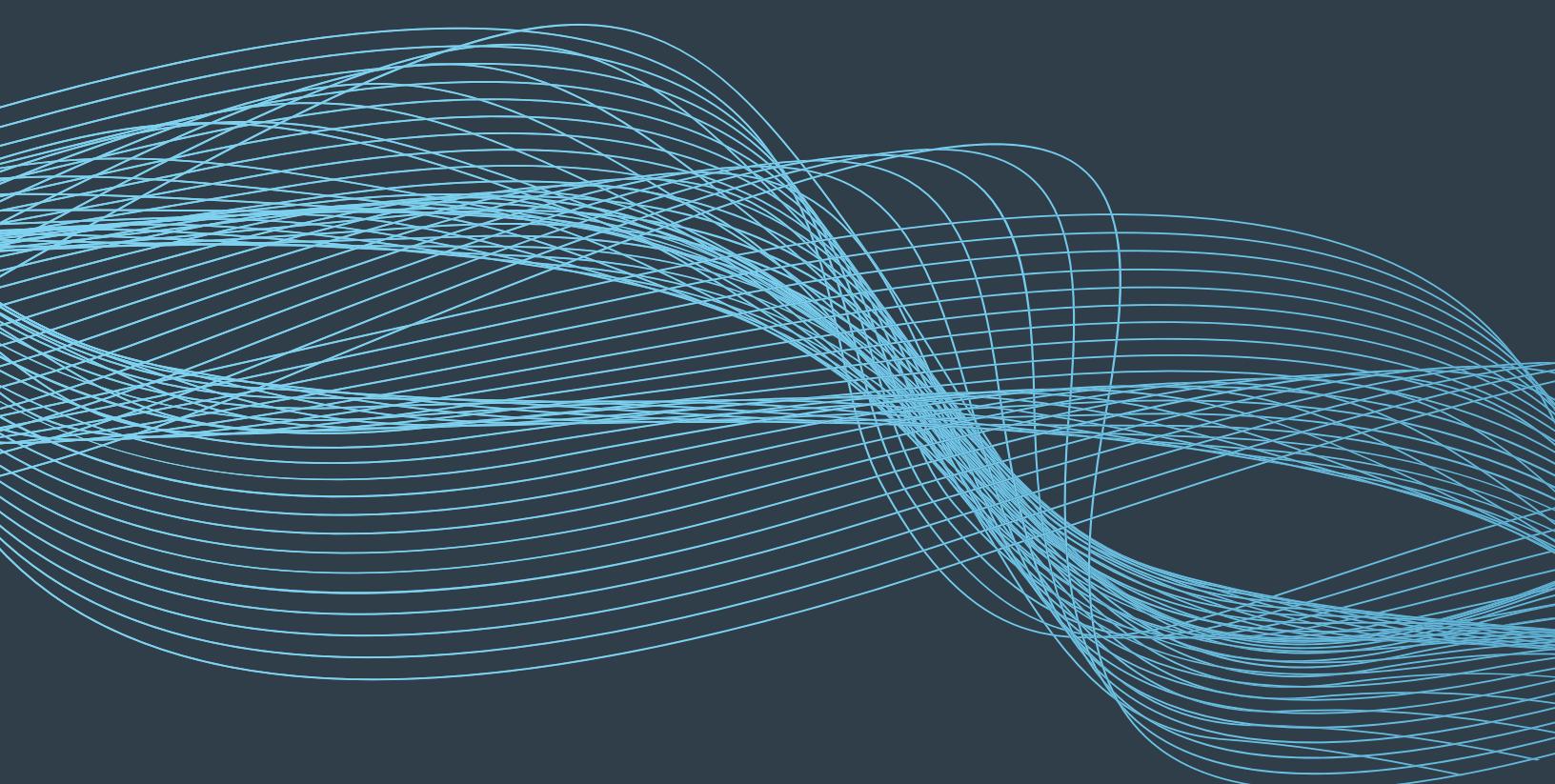
Forte de cette solution, l'institution a réduit le risque de non-conformité et les coûts de conformité en consacrant moins de temps et de ressources au reporting, tout en améliorant la confiance dans les données sur les risques. L'extrême évolutivité du cloud, des technologies de data lake et de la solution Talend a permis de recueillir et de traiter chaque jour des milliards de lignes de données provenant de sources hétérogènes.

La solution garantit une exactitude et une précision accrues du reporting et une détermination des risques transparente. Elle offre aussi plus de visibilité et des mesures concrètes pour résoudre les problèmes de qualité des données. Le profilage génère des données exploitables car le logiciel identifie les éventuelles failles de données, détermine la cause principale et permet aux acteurs disposant des outils nécessaires de les résoudre.

En associant Talend Platform et AWS, l'institution collecte des données dans son data lake, les rapproche du glossaire métiers des « éléments de données critiques » et retrace en détail l'ensemble du processus d'agrégation des données sur les risques. Talend Data Quality est ensuite utilisé pour établir des rapports en fonction d'indicateurs de la qualité, déclenche des alertes et permet aux data stewards de résoudre les problèmes liés aux données. L'entreprise s'appuie désormais sur des données fiables et partagées pour son reporting métier et son reporting réglementaire sur les risques. La solution de conformité, complétée par des bases technologiques solides et extrêmement évolutives, a été mise en œuvre selon une approche « commencer petit, se développer rapidement » qui a vait été élaborée dans un premier temps pour un cas d'usage particulier, mais peut être étendue à toutes l es activités réglementées.

La gouvernance des données moderne permet aux entreprises de relever les défis posés par la conformité réglementaire

- En capturant des volumes importants, à la croissance exponentielle, de données très diverses provenant de sources multiples
- En créant un inventaire des données pour retracer l'origine des données, leur destination et leur mode de traitement
- En vérifiant la qualité de ces données et en suivant facilement les problèmes de qualité
- En générant des rapports fiables rapidement, sans corrections manuelles
- En favorisant la responsabilisation par rapport aux données, en mettant en place des politiques et des règles de protection des données



Chapitre 7 :

De l'intégration à
l'intégrité des données

Pourquoi vos collaborateurs doivent avoir une bonne maîtrise des données

Maintenant que nous avons une vue d'ensemble de ce qu'il faut faire pour obtenir des données fiables, nous ne devons pas oublier que la gouvernance des données est une entreprise de longue haleine. Elle doit tenir compte du fait que les collaborateurs, les processus, les entreprises et les clients sont à des stades différents de leur courbe de maturité.

C'est là qu'entre en scène l'approche « commencer petit, se développer rapidement » (start small and grow fast) de Talend. La plupart des initiatives orientées données débutent par la création d'un « emplacement de données » où les entreprises capturent toutes leurs données. Quel que soit le nom qu'on lui donne, hub de données, entrepôt de données, data lake ou vue à 360° des clients, la logique est la même.

Tout commence généralement par la capture ou le mouvement de données, suivi par la transformation (par exemple, agrégation ou rapprochement). C'est là que débute la gouvernance des données et que nous intervenons pour construire et gérer les pipelines de données en temps réel. En ce qui concerne la gouvernance des données, c'est l'origine de la gestion des données.

L'impératif de données fiables impose que vous placiez la qualité des données au cœur de votre stratégie de gouvernance des données.

Une fois en place, vous pouvez prendre le contrôle de vos données grâce à des fonctions puissantes de profilage, de correspondance (avec le machine learning) et de masquage des données. Si vous souhaitez analyser en profondeur les défis liés à l'intégrité, nous vous conseillons de télécharger notre [Guide complet de la qualité des données](#).

Vers l'intelligence des données



Ce schéma simplifié vous présente les trois niveaux de maturité pour devenir une entreprise intelligente orientée données.

Ceci dit, la route vers l'intégrité des données est semée d'embûches. L'un des plus grands obstacles auxquels vous vous heurterez est la capacité de vos communautés à comprendre en quoi les données constituent un atout et comment elles peuvent être améliorées.

Selon [Accenture](#), 78 % des dirigeants d'entreprise prévoient que leur entreprise passerait au numérique, pourtant seuls 49 % d'entre eux ont déclaré avoir une stratégie de gestion et de développement des compétences nécessaires dans le monde numérique.

La « dataalphabétisation » : un concept encore confidentiel

D'un côté, les entreprises sont vivement encouragées à investir dans un stockage plus critique et plus flexible leur permettant de s'adapter à un volume croissant de données. De l'autre, elles placent des outils décisionnels et collaboratifs à la disposition de leurs collaborateurs.

En mettant l'accent sur les outils numériques, elles oublient souvent de consacrer du temps et des ressources pour favoriser l'élément essentiel pour obtenir des données fiables à grande échelle : la « dataalphabétisation » ou culture des données. Ce principe s'applique également à votre organisation, au niveau des entreprises.

Une datalphabétisation insuffisante peut mettre en péril la transformation de l'intégrité de vos données

Bien que les processus et les responsabilités soient d'une importance primordiale, il est essentiel de mettre l'accent sur les capacités et les connaissances de vos collaborateurs afin qu'ils puissent contribuer de façon active à vos programmes de données. Dans le cas contraire, il est très probable que vos efforts soient vains.

Avant de déployer votre stratégie de données, mesurez les différents niveaux d'expertise sur une échelle allant de « expertise supérieure aux fondamentaux » à « expertise avancée ». Il est essentiel d'envisager d'évaluer les capacités de vos collaborateurs afin de pouvoir adapter en conséquence votre programme d'apprentissage. Faute de quoi, ce dernier risquera d'être incohérent ou inadapté et n'intéressera pas vos collaborateurs, voire entravera la diffusion de votre stratégie de données.

Votre programme d'apprentissage doit être soumis à des étapes essentielles si vous voulez améliorer le niveau d'intégrité

Dans l'idéal, vous rencontrerez les membres des départements Ressources humaines et Formation et développement pour leur expliquer vos intentions et faire en sorte que l'offre en matière de développement professionnel de votre entreprise corresponde à vos besoins. C'est une excellente occasion de demander une analyse de l'apprentissage et d'évaluer les compétences numériques disponibles au sein des départements qui font partie intégrante de votre stratégie de données.

D'ici 2020, 80 % des entreprises mettront en place un développement volontaire des compétences dans le domaine de la datalphabétisation.

À retenir :

La datalphabétisation est la capacité à lire, écrire et communiquer des données en contexte, y compris la compréhension des sources de données et des concepts, des méthodes et des techniques analytiques appliquées, et la capacité à décrire le cas d'usage, l'application et la valeur résultante. Ou de façon plus informelle, « [parlez-vous data ?](#) »

Vous pouvez utiliser la contrainte réglementaire du RGPD pour rendre cette formation obligatoire au sein de votre entreprise. Pour vous faire une idée des types de programmes d'apprentissage disponibles, opérez une sélection dans LinkedIn pour en voir le contenu. Si vous êtes intéressé par des spécialistes des données dans certains domaines d'activité (commerciaux par exemple) envisagez avec le département Formation et développement d'investir dans une application mobile d'apprentissage, afin que vous puissiez les cibler, au moment où vous en avez besoin, sur leur téléphone portable, où qu'ils soient. Assurez-vous que le département Formation et développement investira dans les bons outils afin d'optimiser l'efficacité de vos programmes de datalphabétisation.

Appliquez le modèle 70-20-10 à votre stratégie de datalphabétisation grâce à des outils numériques

Au moment de définir votre programme de compétences, veillez à choisir des applications numériques inspirées du modèle [70-20-10](#) qui utilise à la fois les interactions sociales, l'expérience de terrain, les discussions partagées et les programmes de formation en ligne/hors-ligne. La clé du succès : obtenir la meilleure participation possible de la part de vos communautés de données :

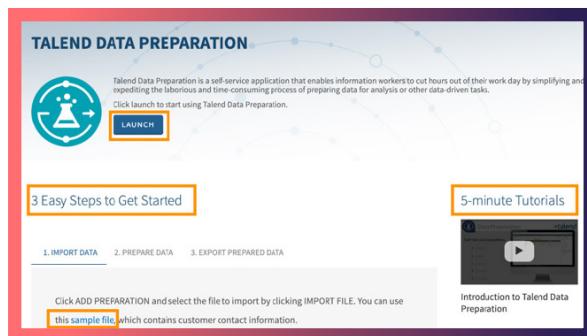
Le modèle 70-20-10 ([70-20-10 Model for Learning and Development](#) en anglais) est un modèle d'apprentissage et de développement professionnel basé sur une répartition proportionnelle des modes d'apprentissage efficaces.

- 70 % — missions intéressantes
- 20 % — relations favorisant le développement
- 10 % — cours et formations

Talend Cloud met à votre disposition tout ce dont vous avez besoin pour lancer un programme de gouvernance des données : versions d'essai, vidéos pratiques étape par étape intégrées et mode d'emploi intégré pour en savoir plus.

Optez pour un apprentissage au sein d'outils en libre-service pour atteindre le niveau d'intégrité voulu.

Envisagez d'investir dans des outils en libre-service à la fois simples et puissants, s'appuyant sur une expérience unifiée, mais veillez à ce que l'apprentissage puisse être déployé par le département IT. La collaboration entre les métiers et l'IT en sera facilitée et le travail nécessaire pour préparer, compiler et protéger les données réduit. Vous pourrez ainsi commencer à utiliser ces outils en libre-service sans qu'une expertise considérable dans la formation soit nécessaire.



Envisagez le déploiement de programmes d'apprentissage mixtes

Les programmes d'apprentissage mixtes doivent être privilégiés pour un apprentissage à grande échelle : il s'agit de parcours structurés au sein d'un système de gestion de l'apprentissage alliant apprentissages en ligne et hors-ligne. Ils seront les plus adaptés au déploiement de vos programmes de dataalphabétisation. Vos collaborateurs se réuniront en ligne et hors ligne dans des salles de formation physiques et virtuelles afin de partager leurs expériences et profiter de l'effet de groupe. Si les administrateurs de vos départements RH et Formation et développement leur déléguent l'autorité, les formateurs pourront également assurer le suivi du programme.

Les programmes d'apprentissage mixtes doivent toujours être privilégiés par rapport aux programmes de formation intégralement en ligne, car la participation à ces derniers est généralement très faible.

Encouragez les membres les plus actifs à développer ces communautés par le biais d'applications d'apprentissage.

The screenshot shows the LinkedIn Learning platform. At the top, there's a navigation bar with 'LEARNING' and 'Library' tabs, a search bar, and user status indicators for 'Home', 'In Progress', and 'Saved'. Below the header, there's a large image of a person holding a key next to a padlock. The main content area has a title 'Technology: Network and System Administration' and a course title 'Become a GDPR and Data Privacy Expert'. A descriptive paragraph about data privacy is followed by three learning objectives: 'Learn why data privacy is a crucial issue, and how to manage it.', 'Identify the global rights and responsibilities of data privacy and governance.', and 'Understand how GDPR affects you in any part of your job that connects to using data.' On the right side, there's a 'Share' button and a 'Learning path details' sidebar showing '8 hours of expert-created content' and '6 items of learning content', with a prominent 'Start learning' button.

Coordonnez vos formations internes avec le déploiement de votre stratégie de données

Assurez-vous que votre programme de formation interne soit synchronisé avec votre programme de gouvernance des données : votre programme de formation doit coïncider avec le calendrier de votre stratégie de données afin que vos collaborateurs comprennent son intérêt et saisissent cette opportunité.

Organisez régulièrement des réunions avec vos collègues du département Formation et développement afin de mesurer régulièrement l'avancement de la formation et décider ensemble des prochaines étapes à mettre en place pour la datalphabétisation.

Par exemple, dans le cadre de la législation européenne sur la protection des données (RGPD), des contenus prêts à l'emploi sont fournis par les entreprises formatrices pour expliquer les principes fondamentaux de la protection des données afin qu'à votre tour, vous puissiez informer vos collaborateurs de ces principes, lors du déploiement des règles de protection des données.

Élaborez un contenu mettant l'accent sur les données en tant que valeur

Assurez-vous que votre formation présente les données non seulement comme un atout précieux à protéger, mais aussi comme une valeur à monétiser. Il arrive souvent que des formations obligatoires mettent l'accent sur les risques à minimiser et non sur les opportunités à exploiter. Il est essentiel que vous donnez à vos collaborateurs les moyens de protéger les données ; mais aussi de compiler et de les valider afin qu'ils se les approprient. Se contenter de proposer une formation sur la protection des données ne suffit pas.

Les données au 21e siècle sont comme le pétrole au 18e : « une ressource précieuse très peu exploitée. Comme avec le pétrole, ceux qui comprennent la valeur fondamentale des données et apprennent à les extraire et à les utiliser récolteront des bénéfices considérables. »

Joris Toonders, WIRED

Chapitre 8 :

**Devenir une entreprise
“ data intelligente ”**

Mettez en œuvre une stratégie de data intelligence durable avec l'aide d'experts

La data intelligence va encore plus loin avec Talend Data Catalog. Cet outil vous permet d'appliquer une approche systématique et automatisée pour documenter votre environnement de données, y créer une source de gouvernance unique et permettre l'accès à des données fiables à l'aide d'une interface de recherche. C'est ce niveau de maturité que vous devez viser à long terme.

Pour y parvenir, il est toujours préférable de se faire aider et d'avoir des conseils externes afin de réaliser rapidement la mise en œuvre avec un partenaire doté des capacités voulues.

Par exemple, si votre programme est motivé par la nécessité de vous conformer à des règlements en matière de protection des données tels que le RGPD, vous devrez être épaulé pour garantir la conformité de vos règles de traitement des données, des personnes, des responsabilités, etc.

Une expertise dans le lignage et le catalogage des données vous permettra de devancer toute exigence de conformité ou piste d'audit ; en effet, vous connaîtrez la provenance de vos données, leurs utilisateurs et leurs liens avec d'autres données.

Il est toujours bénéfique d'avoir un partenaire externe si vous souhaitez devenir rapidement une entreprise « data intelligente » et bénéficier de ses enseignements sur d'autres projets de gouvernance. Ce partenaire repérera immédiatement les difficultés et identifiera avec vous les solutions à adopter pour éliminer les obstacles. Vous gagnerez en vitesse et en expérience, et minimiserez les risques d'échec en mettant en place la bonne approche.

Veillez à ce que votre partenaire ne se contente pas de tirer les conclusions qui s'imposent et de réaliser une intégration simple, mais puissante. Dans le cadre de la gouvernance des données, il doit disposer de compétences en catalogage et en lignage des données, ainsi qu'en méthodologie, en conseil et en gestion des données. Votre partenaire doit réunir des compétences techniques et des services de conseil.

Une fois encore, le choix du partenaire adéquat signifie également s'entourer de consultants adaptés à votre entreprise. Et surtout, poursuivez cette collaboration. Nous sommes souvent témoins de l'échec de projets de gouvernance en raison d'un manque de suivi ou de direction au fil du temps. Il vous faut donc choisir un partenaire offrant une approche de conseil, de bonnes connaissances de la gestion des données et un personnel compétent de premier ordre qui restera à vos côtés, du début à la fin du projet.

Voyez votre projet de gouvernance des données comme la construction de la maison de vos rêves. Vous aurez besoin d'un plan et d'un architecte pour vous guider tout au long de la construction, avec les ressources adaptées.

Sinon, vous risquez de passer beaucoup trop de temps avec différents sous-traitants et de dépasser votre budget.

Élaborez votre cadre de gouvernance des données dès aujourd’hui

Talend (NASDAQ : TLND), leader des solutions d'intégration cloud, libère les données des infrastructures existantes pour mettre plus rapidement de données pertinentes au service de votre entreprise. La plate-forme Talend Cloud assure l'intégration des données dans les environnements sur site et cloud (public, privé ou hybride) et permet une meilleure collaboration entre les équipes IT et les équipes métiers. Son architecture native, ouverte et extensible lui permet de s'adapter rapidement aux innovations et de répondre à moindre coût aux demandes sans cesse croissantes des volumes de données, des utilisateurs et des cas d'usage.

Plus de 1 500 entreprises du monde entier ont confié leurs données à Talend, y compris GE, HP Inc. et Domino's. Talend a été reconnu comme leader dans son domaine par les principaux cabinets d'analyse et par les publications du secteur, y compris Forbes, InfoWorld et SD Times.

Pour plus d'informations, rendez-vous sur le site fr.talend.com.



talend