



**Les 10 causes principales
des problèmes de qualité de données**

White Paper

Sommaire

Erreurs typographiques et données non conformes.....	3
Obfuscation d'informations	4
Informatique traîtresse et fichiers Excel	6
Après la fusion.....	7
Le changement est bon... Sauf pour la qualité de données	8
Code caché.....	10
Transition des transactions	11
Métamorphose de métadonnées	12
Définition de la qualité de données	13
Perte d'Expertise	14
Conclusion	15
A propos de Talend	16

Nous reconnaissons tous les problèmes de qualité de données lorsque nous en voyons. Ils peuvent ébranler la capacité de votre entreprise à travailler efficacement, à respecter les réglementations gouvernementales et à faire des bénéfices. Les problèmes techniques spécifiques comprennent les données manquantes, les attributs saisis dans le mauvais champ, les enregistrements en doublon et les modèles de données cassés, pour n'en citer qu'une partie.

Mais plutôt que de rafistoler de mauvaises données, la plupart des experts pensent que la meilleure stratégie pour lutter contre les problèmes de qualité de données est de comprendre les causes à la racine et d'implémenter de nouveaux processus afin de les éviter. Ce livre blanc traite des dix causes principales des problèmes de qualité de données et suggère des implémentations à effectuer dans votre entreprise afin de les prévenir.



Erreurs typographiques et données non conformes

Malgré l'automatisation quasi systématique de l'architecture des données, des informations sont toujours saisies dans des formulaires Web et d'autres interfaces utilisées par les clients. Les erreurs de saisie sont une source fréquente d'imprécision des données. Les gens font parfois des fautes de frappe. Ils choisissent la mauvaise entrée dans une liste. Ils saisissent la bonne donnée au mauvais endroit.

Etant donnée la complète liberté laissée pour renseigner un champ, les personnes saisissent des données de mémoire. Le nom du vendeur est-il Grainger, WW Granger, ou W. W. Grainger ? Idéalement, il devrait y avoir un ensemble de données de référence d'entreprise (métadonnées) afin que les formulaires permettent aux utilisateurs de trouver les bons vendeurs, noms de clients, villes, références, etc.

Stratégie de l'entreprise

- Formation - Assurer que les personnes saisissant les données connaissent leur impact sur les applications en aval.
- Définition des métadonnées - En restreignant ce que les gens peuvent saisir dans un champ via une liste exhaustive, de nombreux problèmes peuvent être évités. Ces métadonnées (pour le nom des vendeurs, les références, etc.) peuvent devenir une partie de la qualité des données dans l'intégration de données, les applications métier et d'autres solutions.
- Monitoring - Rendre publics le résultat des données mal saisies et féliciter les personnes ayant saisi des données correctement. Vous pouvez en garder un suivi grâce à un logiciel de monitoring de données, tel que Talend Data Quality Portal.
- Validation en temps réel - En plus des formulaires, des outils de validation de qualité de données peuvent être implémentés afin de valider des adresses, des adresses e-mail ainsi que d'autres informations importantes lors de la saisie. Assurez-vous que votre solution de qualité de données puisse diffuser la qualité de données dans des environnements de serveurs d'application, dans le cloud ou dans un Enterprise Service Bus (ESB).

#2

Obfuscation d'informations

Les erreurs lors de la saisie de données ne sont pas toujours faites par erreur. Combien de fois les gens ne donnent-ils pas des informations incomplètes ou incorrectes afin de protéger leur vie privée ? Si cela n'a aucune incidence sur les personnes saisissant les données, elles ont tendance à les falsifier.

Même si les personnes saisissant les données souhaitent le faire correctement, c'est parfois impossible. Si un champ n'est pas disponible, un autre champ est souvent utilisé. Cela peut causer des problèmes de qualité de données, tels que des numéros de TVA dans le champ du nom, ou des coordonnées dans le champ des commentaires.

Stratégie de l'entreprise

- Récompense - Offrir une prime aux gens saisissant des données personnelles correctement. Cela doit être centré sur les gens saisissant des données de l'extérieur, par exemple via les formulaires Web. Les employés ne doivent pas avoir besoin d'une récompense pour faire leur travail. Le type de récompense dépendra de l'importance d'avoir les informations correctes.
- Accessibilité - En tant qu'expert en charge de l'arbitrage de données, soyez ouvert et acceptez les critiques des utilisateurs. Soyez à l'écoute lorsque les changements de processus nécessitent un changement de technologie. Si vous n'êtes pas accessible, les utilisateurs chercheront tous les moyens de valider leur formulaire, même incorrect.
- Validation en temps réel - En plus des formulaires, des outils de validation de qualité de données peuvent être implémentés afin de valider des adresses, des adresses e-mail ainsi que d'autres informations importantes lors de la saisie.

#3

Informatique traîtresse et fichiers Excel

Un renégat est une personne qui déserte et trahit un ensemble de principes d'une organisation. C'est exactement ce que font sans le savoir certains chefs d'entreprise impatientes en déplaçant leurs données dans et hors de leurs solutions métier, bases de données, etc. Plutôt que d'attendre de l'aide d'équipes informatiques professionnelles, des équipes métier zélées peuvent décider de créer leur propre ensemble d'applications locales, sans connaissance informatique particulière. Si l'application doit répondre aux besoins immédiats du département, il est peu probable qu'elle soit conforme aux standards de données, de modèles de données ou d'interfaces. La base de données doit commencer par faire une copie d'une base de données approuvée vers une application locale sur le bureau des membres de l'équipe. D'importants morceaux de données stockés dans des feuilles de calcul Excel, dénommés « spreadmarts », sont facilement duplicables sur le bureau des membres de l'équipe. Dans ce scénario, vous perdez le contrôle des versions et des standards. Il n'y a pas de sauvegarde, de versionnement ou de règle métier.

Stratégie de l'entreprise

- Culture d'entreprise - Il devrait y avoir des conséquences pour les personnes propageant ces spreadmarts, les dissuadant de créer des applications de données locales.
- Communication - Eduquez et formez vos employés sur les conséquences négatives des spreadmarts.
- Gestion de « Small data » - Il est crucial d'avoir des outils permettant à des utilisateurs métier et aux professionnels informatique de gérer les données. Les solutions comme Talend Master Data Management (MDM) permettent de combler les

lacunes entre les applications informatiques coûteuses et limitées et la gestion métier effective des données.

- Verrouiller les données - Le but est d'atteindre une culture d'entreprise où créer des spreadsheets sans sanction n'est pas possible. Certaines entreprises ont trouvé le succès en verrouillant les données afin de les rendre plus difficiles à exporter.

#4

Après la fusion

Les fusions d'entreprises augmentent le risque d'erreurs de qualité de données car elles se déroulent en général rapidement et ne sont pas prévues par les départements informatiques. Presque immédiatement, il y a une certaine pression pour consolider et raccourcir le planning. La consolidation inclura probablement le besoin de partager des données dans un ensemble varié d'applications disjointes. De nombreux raccourcis sont pris pour rendre cela possible, ce qui comprend souvent des risques connus ou inconnus pour la qualité de données.

Parmi les priorités du planning raccourci, la fusion des départements informatiques peut provoquer un conflit culturel et différentes versions de la vérité. De plus, des fusions peuvent créer une perte d'expertise lorsque des personnes clés partent au milieu du projet pour chercher de nouvelles aventures.

Stratégie de l'entreprise

- Conscience d'entreprise - Lorsque c'est possible, la direction doit ordonner la division du travail afin d'éviter des conflits de culture et la rétention de l'information par les gens assoiffés de pouvoir.

- Document - Votre projet informatique doit perdurer, même si l'équipe entière part, se disperse ou est écrasée par un bus en traversant la rue. Vous pouvez réussir à le maintenir avec une bonne documentation de l'infrastructure.
- Consultants externes - Le management doit savoir qu'il y a du travail supplémentaire à faire et que des conflits peuvent émerger après une fusion. Des consultants peuvent fournir la continuité nécessaire pour réussir la transition.
- Gestion de données agile - Les solutions open source de data management Talend permettent de garder votre entreprise flexible, vous donnant ainsi la possibilité de diviser et conquérir la charge de travail sans licence coûteuse d'applications commerciales.

#5

Le changement est bon... Sauf pour la qualité de données

Les entreprises subissent des changements de leurs processus métier afin de s'améliorer. C'est une bonne chose, n'est-ce pas ? Les principaux exemples comprennent :

- Expansion de l'entreprise sur de nouveaux marchés
- Nouveaux accords de partenariats
- Nouvelles lois de régulation sur le reporting
- Reporting financier à une entreprise mère
- Réduction d'effectifs

Si la qualité de données est définie comme l'adaptation aux besoins (« fitness for purpose »), que se passe-t-il lorsque l'objectif change ?

C'est cette nouvelle utilisation de données qui apporte des changements dans le niveau perçu de qualité de données, même si les données sous-jacentes sont les mêmes. Il est naturel que les données changent. Lorsqu'elles le font, les règles de qualité de données, les règles métier et les couches d'intégration de données doivent également changer.

Stratégie de l'entreprise

- Gouvernance des données - En mettant en place une équipe de gouvernance de données transverse aux métiers, vous aurez toujours une équipe qui contrôlera les changements que subit votre entreprise et examinera leur impact sur les informations. Cela devrait être dans la charte d'une équipe de gouvernance de données.
- Communication - Une communication régulière et un modèle de métadonnées bien documenté simplifient le changement.
- Flexibilité des outils - L'un des défis lorsque vous achetez des outils de qualité de données embarqués dans des applications d'entreprise est qu'ils peuvent ne pas fonctionner dans toutes les applications d'entreprise. Lorsque vous choisissez des outils, soyez sûr qu'ils soient suffisamment flexibles pour fonctionner avec les données de n'importe quelle application et que l'entreprise soit attachée à la flexibilité et à l'ouverture.

#6

Code caché

Les bases de données commencent rarement leur vie en étant vides. Le point de départ est généralement une conversion de données à partir d'une source de données déjà existante. Le problème étant que, si les données peuvent fonctionner parfaitement dans l'application source, elles peuvent échouer dans la cible. Il est difficile de voir tout le code applicatif maison et les processus spéciaux qui se déroulent derrière les données, sauf si vous les profilez.

Stratégie de l'entreprise

- Profilez tôt et souvent - Ne partez pas du principe que vos données sont adaptées à l'objectif visé parce qu'elles fonctionnent dans l'application source. Le profiling vous donnera une évaluation exacte de la forme et de la syntaxe des données dans la source. Il vous permettra également de savoir la quantité de travail que vous devrez fournir pour les faire fonctionner dans la cible.
- Appliquez des outils de qualité de données lorsque c'est possible
 - Plutôt que du code personnalisé dans l'application, une meilleure stratégie est de laisser les outils de qualité de données appliquer les standards. Les outils de qualité de données appliquent des standards d'entreprise de manière uniforme, ce qui simplifie le partage des données.



Transition des transactions

De plus en plus de données sont échangées entre les systèmes via des interfaces en temps réel (ou quasiment en temps réel). Dès que les données entrent dans une base de données, cela déclenche des procédures nécessaires à l'envoi de transactions à d'autres bases de données en aval. L'avantage est la propagation immédiate des données à toutes les bases de données correspondantes.

Mais que se passe-t-il lorsque les transactions tournent mal ? Un système qui fonctionne mal peut causer des problèmes aux applications en aval. En fait, même un petit changement dans un modèle de données peut poser des problèmes.



Stratégie de l'entreprise

- Vérifications des schémas - Effectuez des vérifications de schémas dans les flux de vos Jobs afin de vous assurer que vos applications en temps réel produisent des données cohérentes. Les vérifications de schémas vont effectuer des tests simples afin de vérifier que vos données sont complètes et correctement formatées, avant de les charger.
- Monitoring de données en temps réel - Le niveau supérieur aux vérifications de schémas est le monitoring de données proactif avec des outils de profiling et de monitoring de données. Les outils comme Talend Data Quality Portal vous assurent que vos données contiennent le bon type d'informations. Par exemple, si vos références ont toujours une forme et une longueur particulières et qu'elles contiennent un ensemble fini de valeurs, toute variation de cet attribut peut être monitorée. Lorsque des variations se produisent, le logiciel de monitoring vous envoie une notification.

#8

Métamorphose de métadonnées

Le référentiel des métadonnées doit pouvoir être partagé entre de nombreux projets, avec un audit régulier sur l'utilisation et l'accès au référentiel. Par exemple, votre entreprise peut avoir des références et des descriptions qui sont universelles pour le CRM, pour la facturation, pour les systèmes ERP, etc. Lorsqu'une référence devient obsolète dans le système ERP, le système CRM doit le savoir. Les métadonnées changent et doivent être partagées.

En théorie, documenter en totalité ce qu'il se passe dans la base de données et combien sont interdépendants les nombreux processus vous permettrait d'atténuer complètement le problème. Les descriptions et les références doivent être partagées entre toutes les applications concernées. Pour commencer, vous pouvez analyser les implications relatives à la qualité de données de tout changement dans le code, les processus, la structure des données, ou les procédures de collection des données, ce qui vous permet d'éliminer les erreurs de données inattendues. En pratique, cette tâche représente un énorme travail.

Stratégie de l'entreprise

- Modèles de données prédéfinis - De nombreuses entreprises possèdent des définitions très simples du contenu de chaque ensemble de données. Par exemple, l'industrie automobile suit certains standards ISO 8000. L'industrie de l'énergie suit les standards Petroleum Industry Data Exchange (PIDX). Cherchez un modèle de données dans votre domaine pour vous aider.
- Gestion de données agile - La réussite de la gouvernance de données se fait en commençant petit et en construisant un processus qui règle d'abord les problèmes les plus importants du point de vue métier. Vous pouvez tirer parti des solutions agiles

de Talend afin de partager des métadonnées et de mettre en place des processus facultatifs à travers l'entreprise.

#9

Définition de la qualité de données

De plus en plus d'entreprises reconnaissent le besoin de qualité de données, mais il y a différents moyens de nettoyer des données et d'améliorer leur qualité. Vous pouvez :

- Ecrire du code et nettoyer manuellement
- Manipuler la qualité de données dans l'application source
- Acheter des outils pour nettoyer les données

Considérez ce qu'il se passe lorsque vous avez au moins deux de ces types de processus de qualité de données ajustant les données. L'équipe commerciale a une définition du client, la comptabilité en a une autre. A cause des différents processus, les deux équipes ne s'accordent pas sur le fait que les deux enregistrements sont des doublons.

Stratégie de l'entreprise

- Outils de standardisation - Choisissez des outils qui ne sont pas liés à une solution particulière. De la qualité de données uniquement dans SAP, par exemple, ne va pas améliorer vos ensembles de données Oracle, Salesforce et MySQL. Lorsque vous sélectionnez une solution, prenez-en une comme Talend Data Quality, capable d'accéder à toutes les données, n'importe où et n'importe quand. Comme la solution de Talend est flexible, bénéficier d'une solution commune comme celle-ci, à travers différentes plateformes et solutions ne sera pas très coûteux.

- Gouvernance de données - En mettant en place une équipe de gouvernance de données transverse, vous avez les bonnes personnes au bon endroit afin de définir un modèle de données commun.

#10

Perte d'Expertise

Dans presque tous les projets ayant de très grands volumes de données, il y a une personne dont l'expertise concernant les données héritées est exceptionnelle. Ce sont ces gens qui comprennent pourquoi la date d'embauche de certains employés est stockée dans le champ de la date de naissance et pourquoi certains attributs de noms contiennent également des numéros de TVA.

Les données peuvent être une sorte d'historique pour une entreprise. Elles peuvent provenir de systèmes hérités. Dans certains cas, la même valeur dans le même champ signifie quelque chose de totalement différent dans un autre enregistrement. Avoir conscience de ces anomalies permet aux experts d'utiliser les données correctement.

Si vous vous trouvez dans cette situation, voici quelques processus métier que vous pouvez suivre.

Stratégie de l'entreprise

- Profiler et monitorer - Profiler les données vous permet d'identifier la plupart de ces types de problèmes. Par exemple, si un numéro de TVA se trouve dans un champ de nom, l'analyse vous permet de le remarquer rapidement. Le monitoring prévient les récurrences.
- Documenter - Même s'ils peuvent être réticents par peur de perdre la sécurité de leur emploi, assurez-vous que les experts

documentent toutes les anomalies et les transformations devant advenir à chaque déplacement des données.

- Utiliser les consultants - Les employés experts peuvent être si précieux et si occupés qu'ils n'ont pas le temps de documenter les anomalies héritées. Les entreprises de consulting externes sont généralement très bonnes pour documenter les problèmes et fournir une continuité entre l'héritage et les nouveaux employés.

Conclusion

Aujourd'hui, la plupart des entreprises ont compris que leur succès est étroitement lié à la qualité de leurs informations. Les entreprises comptent sur les données pour prendre des décisions importantes pouvant affecter le service clients, la conformité aux réglementations, la chaîne d'approvisionnement et bien d'autres domaines encore. Puisque vous collectez de plus en plus d'informations concernant les clients, les produits, les fournisseurs, les transactions et la facturation, vous devez vous attaquer à la racine des problèmes de la qualité de données. Les outils de qualité de données sont simplement cela, des outils pour attaquer les causes à la racine et résoudre les problèmes n'ayant pas été détectés à la racine. Les outils vont de pair avec les gens et les processus pour conduire à une information de haute qualité et de haute valeur ajoutée.

Vous souhaitez en savoir plus à propos de la suite open source Talend Data Quality ? Visualisez les Webinars en ligne ou téléchargez la dernière version du logiciel open source de data management sur Talend.com.

A propos de Talend

Talend est le leader reconnu du marché en data management open source. Un studio de développement unique fournit une cohérence entre les projets d'intégration, de qualité et de gestion de données, afin que les ressources puissent être partagées et réutilisées, tout en restant ouvertes, intuitives and économiques. Relever les défis de la gestion de données ne signifie pas être excessivement cher ou restreint à une application. Talend bouleverse le modèle propriétaire traditionnel en fournissant une technologie open source établie sur la base des performances, de la simplicité d'utilisation, de l'extensibilité et de la robustesse.

Les produits de qualité de données de Talend fournissent aux entreprises une vue et un monitoring détaillés des données métier et incluent un ensemble complet de fonctionnalités permettant d'améliorer la qualité et l'efficacité des actifs de données critiques.

Talend propose deux options de qualité de données : Talend Open Profiler, un produit de profiling de données open source, disponible sur le site Web de Talend en téléchargement gratuit et Talend Data Quality, qui comprend des fonctionnalités supplémentaires de niveau Entreprise, notamment le nettoyage et la standardisation, le matching et le dédoublement, un outil embarqué d'intégration de données pour effectuer rapidement et facilement des transformations de données, et des dashboards de qualité de données basés Web.