Chapitre 3
Préparer vos données pour en exploiter le potentiel

1. Qualité des données : rappel

La qualité des données est un élément fondamental à considérer avant d'aborder les techniques de nettoyage et de traitement des données. Pour toute organisation cherchant à prendre des décisions éclairées et à tirer le meilleur parti de ses informations, cet aspect ne peut être négligé. Toutes les procédures que nous examinerons dans ce chapitre ont un seul but : mettre à la disposition des différentes équipes des données fiables !

1.1 Qu'est-ce que la qualité des données ?

La qualité des données reflète la capacité d'une organisation à maintenir l'exactitude et la pérennité de ses informations au cours du temps. En tant qu'experts du domaine, nous devons fournir des données irréprochables tout en nous appuyant sur des indicateurs clairs et facilement interprétables. Nous commencerons par examiner en détail les six critères qui définissent la qualité des données (QDD).

Cette notion englobe à la fois les caractéristiques intrinsèques des données et les méthodes mises en œuvre pour les garantir. En essence, la qualité des données se définit par leur aptitude à servir l'usage auquel elles sont destinées.

Une initiative de qualité des données s'inscrit dans la durée et s'intègre à l'ensemble du cycle de vie des données. Elle requiert une évolution culturelle dans la façon dont l'organisation gère ses données. C'est une approche globale qui impacte l'ensemble de l'entreprise et ses pratiques quotidiennes.

Il est important de noter que des données erronées en entrée d'un processus produiront inévitablement des résultats inexacts en sortie. Par conséquent, une stratégie fondée sur des données de piètre qualité aboutira à des décisions inefficaces, avec des conséquences directes sur le retour sur investissement.

Crédits : https://dataedo.com/

© Editions ENI - All rights reserved

1.2 Pourquoi est-ce que la QDD est importante ?

La qualité des données est souvent compromise par différents facteurs. On peut citer par exemple les erreurs humaines lors de la saisie initiale. En effet, les fautes de frappe, les conventions de nommage différentes entre les sources de données ou des abréviations incorrectes sont une source fréquente de problèmes. De plus, les informations initialement exactes peuvent devenir obsolètes au fil du temps en raison de l'évolution du contexte.

La qualité des données est souvent compromise par divers facteurs, entraînant des conséquences coûteuses pour les entreprises. Voici quelques exemples concrets de problèmes fréquemment rencontrés et leurs impacts :

– Erreurs de saisie : en 2018, un employé de Samsung Securities a commis une erreur de saisie monumentale lors de la distribution de dividendes, émettant par inadvertance 2,8 milliards d'actions « fantômes », soit environ trente fois le nombre total d'actions existantes. Cette erreur a provoqué une perturbation massive du marché et une perte de confiance envers la gestion de l'entreprise, nécessitant des mesures correctives coûteuses. [Source : https://www.sirfull.com/blog/impacts-mauvaise-gestion-donnees/]
– Incohérences de données : en 2022, Unity Technologies a fait face à un problème majeur avec son outil de publicité ciblée Pinpointer. Des données inexactes provenant d'un grand client ont corrompu les modèles d'IA, entraînant une perte de revenus de 110 millions de dollars et une chute de 37 % de la valeur des actions de l'entreprise. [Source : https://zeenea.com/fr/quelles-sont-les-principales-erreurs-liees-a-la-data-quality-et-comment-les-resoudre/]

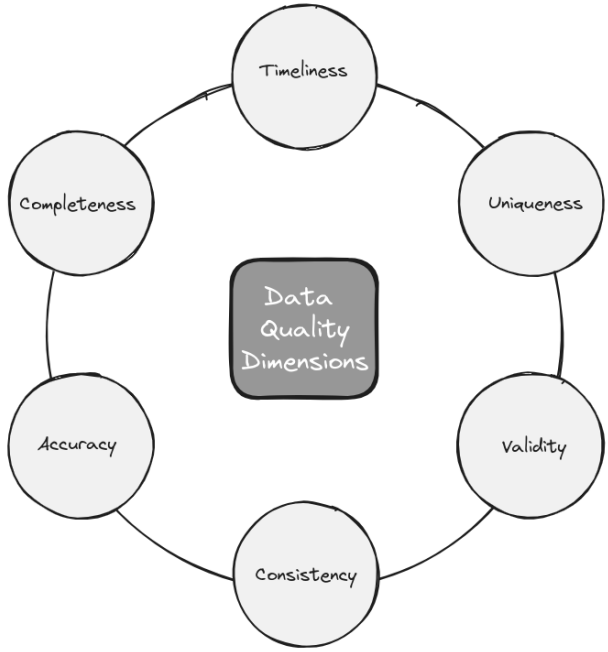
- Données obsolètes : Tesco, le géant britannique de la distribution, a connu des problèmes liés à des erreurs dans ses données de stock, conduisant à des ruptures fréquentes dans ses magasins. Ces inexactitudes ont entraîné une perte de ventes, une insatisfaction des clients et un impact négatif direct sur le chiffre d'affaires de l'entreprise. [Source : https://www.sirfull.com/blog/impacts-mauvaise-gestion-donnees/]
- Erreurs dans les données critiques : entre mars et juillet 2017, Equifax a généré des scores de crédit incorrects pour des millions de consommateurs en raison de données erronées. L'entreprise a dû faire face à des amendes réglementaires, des poursuites judiciaires et une perte de crédibilité, mettant en danger les décisions de prêt et érodant la confiance du public. [Source : https://www.datagalaxy.com/fr/blog/big-data-attention-aux-donnees-de-mauvaise-qualite/]
- Données de localisation inexactes : Royal Dutch Shell a rencontré des erreurs dans les données de localisation de ses puits de pétrole, conduisant à des forages inefficaces. Ces erreurs ont entraîné des coûts supplémentaires de plusieurs millions de dollars et une perte de temps considérable due à des forages incorrects. [Source : https://solutions-business-intelligence.fr/data-quality-enjeux-et-bonnes-pratiques/]

Chacun de ces problèmes de qualité des données a eu des répercussions significatives sur les opérations, la réputation et les résultats financiers des entreprises concernées. Ces exemples soulignent l'importance d'investir dans des processus robustes de gestion de la qualité des données pour éviter ces pièges coûteux.

© Editions ENI - All rights reserved

1.3 Les principaux critères de la QDD

Découvrons les principaux critères de la QDD. Ces critères essentiels sont au cœur de toute stratégie efficace de gestion des données.



1.3.1 Exactitude (accuracy)

L'exactitude est le degré de conformité des données stockées avec les valeurs réelles qu'elles sont censées représenter. Elle mesure la proximité entre la valeur enregistrée et la valeur correcte ou acceptée comme étant vraie.

Dans le secteur bancaire, imaginons une erreur dans le calcul des intérêts d'un prêt immobilier. Si le taux d'intérêt est incorrectement enregistré à 3,5 % au lieu de 3,05 %, cela pourrait entraîner une surfacturation pour des milliers de clients. Non seulement cela pourrait conduire à des pertes financières importantes pour la banque en cas de remboursement, mais cela pourrait aussi gravement nuire à sa réputation et potentiellement entraîner des sanctions réglementaires.

Pour assurer l'exactitude, les entreprises peuvent mettre en place des processus de validation des données, des contrôles croisés automatisés, et des audits réguliers. L'utilisation de l'intelligence artificielle pour détecter les anomalies peut également être efficace.

1.3.2 Exhaustivité (completeness)

L'exhaustivité est la mesure dans laquelle tous les éléments de données nécessaires sont présents dans un ensemble de données spécifique. Elle évalue le degré auquel toutes les valeurs requises sont incluses et tous les enregistrements qui devraient être présents le sont effectivement.

Dans le domaine de la recherche médicale, imaginons une étude sur l'efficacité d'un nouveau traitement contre le cancer. Si les données de suivi post-traitement sont incomplètes pour un groupe significatif de patients, cela pourrait fausser les conclusions de l'étude. Des décisions cruciales concernant l'approbation ou le rejet du traitement pourraient être basées sur des informations partielles, affectant potentiellement la vie de nombreux patients.

Utiliser des champs obligatoires dans les formulaires de saisie, mettre en place des alertes pour les données manquantes, et implémenter des processus de collecte de données systématiques peuvent améliorer l'exhaustivité.

1.3.3 Cohérence (consistency)

La cohérence fait référence à l'absence de contradictions dans les données au sein d'un ensemble de données ou entre différents ensembles de données. Elle assure que les données sont uniformes et logiquement compatibles dans tous les systèmes, applications et processus de l'organisation.

Dans une multinationale, imaginons que le département des ressources humaines et le département finance utilisent des systèmes différents pour gérer les informations des employés. Si un employé change de poste et que cette information n'est mise à jour que dans le système RH, cela pourrait conduire à des erreurs dans la paie, les avantages sociaux, et même dans la planification stratégique des ressources humaines.

© Editions ENI - All rights reserved

L'utilisation d'un système d'information intégré, la mise en place de processus de synchronisation automatique entre différents systèmes, et l'établissement de règles de gestion des données cohérentes à l'échelle de l'entreprise peuvent améliorer la cohérence.

1.3.4 Actualité (timeliness)

L'actualité mesure le degré auquel les données sont à jour et disponibles dans le délai requis pour leur utilisation prévue. Elle évalue la fraîcheur des données par rapport au moment de leur création ou de leur dernière mise à jour, ainsi que leur disponibilité au moment où elles sont nécessaires pour les processus métier.

Dans le domaine du trading haute fréquence, où des décisions d'achat et de vente sont prises en millisecondes, l'actualité des données est critique. Un retard de quelques secondes dans la mise à jour des prix des actions pourrait entraîner des pertes financières considérables. Par exemple, si un événement majeur affecte le cours d'une action, mais que cette information n'est pas reflétée immédiatement dans les données utilisées par les algorithmes de trading, cela pourrait conduire à des décisions d'investissement désastreuses.

L'utilisation de systèmes de traitement en temps réel, la mise en place de processus de mise à jour automatique des données, et l'optimisation des flux de données peuvent améliorer l'actualité.

1.3.5 Validité (validity)

La validité est la mesure dans laquelle les données sont conformes aux règles métier définies, aux formats spécifiés et aux contraintes du domaine. Elle assure que les valeurs des données respectent les critères syntaxiques et sémantiques établis pour le type de données en question.

Dans le secteur de l'aviation, imaginons un système de réservation qui accepte une date de vol antérieure à la date actuelle. Cela pourrait conduire à des problèmes majeurs dans la planification des vols, la gestion des équipages, et potentiellement compromettre la sécurité si ces données invalides sont utilisées dans d'autres systèmes critiques.

L'implémentation de contrôles de validation stricts dans les interfaces de saisie, l'utilisation de contraintes au niveau de la base de données, et la mise en place de processus de nettoyage régulier des données peuvent améliorer la validité.

1.3.6 Unicité (uniqueness)

L'unicité est la propriété selon laquelle chaque entité distincte du monde réel est représentée une et une seule fois dans l'ensemble de données. Elle garantit l'absence de doublons ou de redondances non intentionnelles dans les enregistrements de données.

Dans le secteur de la santé, imaginons un patient ayant plusieurs dossiers médicaux en raison de doublons dans la base de données. Cela pourrait conduire à des erreurs graves dans le traitement si un médecin n'a pas accès à l'historique médical complet du patient. Par exemple, des allergies ou des interactions médicamenteuses pourraient être manquées, mettant en danger la santé du patient.

L'utilisation d'identifiants uniques, la mise en place de processus de déduplication, et l'implémentation de contrôles stricts lors de la création de nouveaux enregistrements peuvent améliorer l'unicité des données.

2. Nettoyage de données

2.1 Premiers pas avec la librairie pandas

Pandas est une bibliothèque Python puissante et polyvalente, conçue pour la manipulation et l'analyse des données. Elle a été développée par Wes McKinney, un chercheur qui a commencé à construire ce qui allait devenir Pandas. Le nom « pandas » est dérivé du terme « Panel Data », un terme d'économétrie pour les jeux de données qui comprennent des observations sur plusieurs périodes.

© Editions ENI - All rights reserved

Pandas est particulièrement adaptée pour travailler avec des données tabulaires, similaires à une feuille de calcul Excel ou une table SQL. Les principales structures de données gérées par cette bibliothèque sont les séries, qui stockent des données selon une dimension, et les DataFrames, qui stockent des données selon deux dimensions (lignes et colonnes). Ces structures de données facilitent la manipulation des données, ainsi que le nettoyage, le prétraitement, l'analyse et la visualisation.

L'utilisation de pandas est largement répandue dans le domaine de l'analyse de données. Elle est souvent présentée comme l'outil idéal pour manipuler des données qui peuvent être organisées sous forme de lignes et de colonnes. De plus, la maîtrise de pandas est une compétence recherchée par les employeurs, car de nombreuses entreprises de tous secteurs utilisent de plus en plus la science des données.

Il existe plusieurs alternatives à pandas, on peut citer notamment polars, dask et cudf. Chacune de ces solutions présente un intérêt, en particulier la vitesse de traitement par rapport à pandas. Nous n'en parlerons pas dans cet ouvrage, car Pandas demeure la bibliothèque la plus utilisée pour l'analyse de données en Python. Sa richesse et sa polyvalence, ainsi que sa large adoption dans la communauté d'analystes de données, en font un outil incontournable.

2.2 Présentation de notre jeu de données

Au cours de ce chapitre, nous allons travailler avec un jeu de données disponible en libre accès sur la plateforme Kaggle.

Le commerce électronique est devenu un nouveau canal pour soutenir le développement des entreprises. Grâce au commerce électronique, les entreprises peuvent accéder à un marché plus large et établir une présence plus importante en fournissant des canaux de distribution moins coûteux et plus efficaces pour leurs produits ou services. Le commerce électronique a également changé la façon dont les gens achètent et consomment des produits et des services. De nombreuses personnes se tournent vers leurs ordinateurs ou leurs appareils intelligents pour commander des biens, qui peuvent être facilement livrés à leur domicile.