Les couturiers inclusifs

Procédure de bonne gestion des données

Amine Ragued
23/07/2024

1. Introduction

La gouvernance des données est un aspect essentiel de la gestion de l'information au sein d'une organisation. Elle vise à garantir que les données soient disponibles, précises, fiables et sécurisées tout au long de leur cycle de vie. La qualité de la gouvernance des données influence directement la capacité d'une entreprise à prendre des décisions informées, à répondre aux exigences réglementaires et à maintenir la confiance des parties prenantes. Cette notice a pour objectif de fournir un cadre structuré pour la gestion des données, en mettant l'accent sur les meilleures pratiques et les techniques clés à chaque étape du processus.

Dans le contexte actuel où les volumes de données augmentent de manière exponentielle et où la réglementation sur la protection des données, telle que le Règlement Général sur la Protection des Données (RGPD), devient de plus en plus stricte, une bonne gouvernance des données est plus cruciale que jamais. Elle permet non seulement de maximiser la valeur des données pour l'entreprise, mais aussi de minimiser les risques associés à leur mauvaise gestion.

Cette notice est divisée en trois sections, chacune abordant un aspect critique de la gouvernance des données :

- **Collecte des Données** : Cette section décrit les méthodes de collecte des données ainsi que les bonnes pratiques à adopter pour s'assurer que les données recueillies sont de haute qualité et conformes aux réglementations.

- **Exploration des Données** : Cette section aborde les techniques d'analyse préliminaire des données, essentielles pour comprendre leur structure et leur contenu, et pour identifier les premières tendances et anomalies.

- **Nettoyage des Données** : Cette section se concentre sur la détection et la correction des anomalies dans les données afin de garantir leur précision et leur fiabilité.

L'objectif ultime de cette notice est de fournir aux gestionnaires de données et aux parties prenantes une feuille de route claire et pratique pour instaurer et maintenir une gouvernance des données efficace. En suivant les principes et les pratiques décrits dans ce document, les organisations peuvent non seulement améliorer la qualité de leurs données mais aussi renforcer leur capacité à tirer parti de ces données pour atteindre leurs objectifs stratégiques.

2. Collecte des Données

La collecte des données est la première étape critique du cycle de vie des données. Elle consiste à acquérir des données à partir de diverses sources de manière structurée et

sécurisée. Une collecte de données bien exécutée garantit que les données obtenues sont de haute qualité, pertinentes et conformes aux exigences réglementaires. Cette section traite des méthodes de collecte des données et des bonnes pratiques à suivre.

2.1 Méthodes de collecte

La collecte des données peut se faire par diverses méthodes, chacune ayant ses propres avantages et défis. Voici quelques exemples de méthodes de collecte des données :

- **Enquêtes et Questionnaires**

Les enquêtes et les questionnaires sont des outils courants pour collecter des données directement auprès des individus. Ils peuvent être administrés en ligne, par téléphone ou en personne. Il est essentiel de concevoir des questionnaires clairs et concis pour obtenir des réponses précises et utiles.

- **Bases de Données Internes**

Les organisations peuvent exploiter leurs propres bases de données internes, telles que les systèmes CRM, ERP ou les archives historiques. Ces sources fournissent souvent des données structurées et standardisées.

- **Sources Externes et Ouvertes**

Il s'agit de l'acquisition de données à partir de sources externes, telles que les bases de données publiques, les ensembles de données gouvernementaux, et les API fournies par des tiers. Ces sources peuvent enrichir les données internes et fournir de nouvelles perspectives.

- **Réseaux Sociaux**

Les réseaux sociaux permettent de collecter des données non structurées provenant de diverses plateformes en ligne. Ces données peuvent offrir des idées précieuses sur les tendances et les opinions des consommateurs.

2.2 Le respect des RGPD

Il est important de s’assurer que la collecte des données respecte le Règlement Général sur la Protection des Données (RGPD).

- **Informer les utilisateurs :** Il est essentiel d’informer clairement les utilisateurs sur la manière dont leurs données seront collectées, utilisées, et protégées. Cela comprend la mise à disposition d’une politique de confidentialité accessible qui détaille :

    - ⊕ Les types de données collectées.
    - ⊕ Les finalités de la collecte de ces données.
    - ⊕ Les méthodes de collecte et d’utilisation des données.
    - ⊕ Les parties avec qui ces données peuvent être partagées.
    - ⊕ Les droits des utilisateurs concernant leurs données.

- **Obtenir leur consentement :** Avant de collecter des données, il est crucial d'obtenir le consentement explicite des utilisateurs. Cela s'applique notamment aux cookies et autres traceurs :

    - ⊕ Utiliser des bannières de cookies claires qui expliquent quelles informations seront collectées et pour quelles finalités.
    - ⊕ Les utilisateurs doivent pouvoir accepter ou refuser les cookies facilement et de manière granulaire (par type de cookies).
    - ⊕ Le consentement doit être documenté et les utilisateurs doivent pouvoir le retirer aussi facilement qu'ils l'ont donné.

- **Collecter uniquement les données nécessaires :** Le principe de minimisation des données impose que seules les données nécessaires à une finalité spécifique soient collectées :

    - ⊕ Déterminer précisément quelles données sont indispensables pour atteindre vos objectifs.
    - ⊕ Éviter de collecter des données excessives ou non pertinentes.
    - ⊕ Réviser régulièrement les données collectées pour s'assurer qu'elles restent pertinentes et nécessaires.

- **Assurer la sécurité des données :** La protection des données personnelles doit être assurée tout au long de leur cycle de vie :

    - ⊕ Utiliser des méthodes de chiffrement pour protéger les données
    - ⊕ Mettre en place des mesures de sécurité physiques et logiques (pare-feu, contrôles d'accès, etc.).
    - ⊕ Former le personnel à la sécurité des données et aux bonnes pratiques de protection.
    - ⊕ Prévoir des procédures de gestion des incidents pour réagir rapidement en cas de violation de données.

- **Respecter les droits des utilisateurs :** Les utilisateurs ont plusieurs droits concernant leurs données personnelles, et il est essentiel de les respecter :

    - ⊕ Droit d'accès : Les utilisateurs peuvent demander à voir les données que vous détenez sur eux.
    - ⊕ Droit de rectification : Les utilisateurs peuvent demander la correction de leurs données inexactes.
    - ⊕ Droit à l'effacement : Les utilisateurs peuvent demander la suppression de leurs données.
    - ⊕ Droit à la limitation du traitement : Les utilisateurs peuvent demander la limitation de l'utilisation de leurs données.

    - ⊕ Droit à la portabilité : Les utilisateurs peuvent demander que leurs données soient transférées à un autre responsable de traitement.
    - ⊕ Droit d'opposition : Les utilisateurs peuvent s'opposer à certains traitements de leurs données.

- **Tenir un registre des activités de traitement :** Les entreprises doivent documenter leurs activités de traitement de données pour prouver leur conformité au RGPD :

- ⊕ Tenir un registre détaillé des traitements incluant les finalités, les catégories de données, les personnes concernées, et les destinataires des données.
- ⊕ Inclure les mesures de sécurité mises en place et les durées de conservation des données.
- ⊕ Mettre à jour ce registre régulièrement et le rendre disponible aux autorités de protection des données sur demande.

En suivant ces points, les organisations peuvent assurer une gestion des données conforme au RGPD, protégeant ainsi les droits des individus et minimisant les risques de non-conformité.

2.3. Bonnes pratiques

Pour garantir que la collecte des données se fait de manière éthique, légale et efficace, il est essentiel de suivre certaines bonnes pratiques :

- **Minimisation des Données :** Collecter uniquement les données nécessaires à l'objectif spécifique de la collecte. Éviter la collecte de données superflues ou excessives qui ne contribuent pas directement aux besoins de l'organisation.

- **Qualité des Données :** Mettre en place des contrôles pour assurer la précision, la cohérence et la complétude des données dès leur collecte. Utiliser des outils de validation et de vérification pour détecter et corriger les erreurs de saisie ou les incohérences.

- **Sécurité des Données :** Assurer la sécurité des données pendant et après la collecte. Utiliser des méthodes de chiffrement et des protocoles sécurisés pour protéger les données contre les accès non autorisés et les fuites.

En suivant ces méthodes et bonnes pratiques, les organisations peuvent s'assurer que la collecte de données est réalisée de manière efficace, éthique et en conformité avec les exigences légales, tout en maximisant la qualité et la fiabilité des données collectées.

3. Exploration des Données

L'exploration des données est une étape cruciale qui permet de comprendre la structure, le contenu et les caractéristiques des données collectées. Elle implique une analyse préliminaire pour identifier des tendances, des anomalies et des patterns. Cette phase initiale est essentielle pour orienter les étapes suivantes de nettoyage, d'analyse et de prise de décision.

Analyse Indépendante des Colonnes

L'analyse indépendante des colonnes constitue une étape fondamentale de l'exploration des données, permettant de se familiariser avec chaque colonne de données de manière isolée. Voici comment procéder :

Identifier le type de données auquel on est confronté

Chaque colonne de données peut contenir différents types de données. Identifier le type de données est essentiel pour choisir les méthodes d'analyse appropriées ; données numériques, données catégorielles, données textuelles, données temporelles

L'identification se fait généralement en examinant les premières lignes de données ou en utilisant des outils de profiling de données.

Comprendre comment ces données sont structurées

La structure des données dans chaque colonne peut varier et nécessite une compréhension approfondie pour une analyse correcte :

Présence de valeurs manquantes : Détecter et quantifier les valeurs manquantes dans chaque colonne. Si ces valeurs sont manquantes, il est possible de les récupérer auprès d'autres bases de données, auprès de collègues ou en identifiant des moyennes afin de compléter la liste des valeurs manquantes.

Variabilité et dispersion : Évaluer la variabilité pour identifier les colonnes avec des valeurs constantes ou très peu variables.

Cette compréhension aide à déterminer les transformations nécessaires et à anticiper les problèmes potentiels dans les étapes ultérieures de traitement.

Identifier les données problématiques

Les données problématiques peuvent inclure des valeurs manquantes, des valeurs aberrantes, des incohérences ou des erreurs de saisie. Identifier ces problèmes est crucial pour garantir la qualité des données :

- ⊕ **Valeurs manquantes** : Utiliser des méthodes de comptage pour identifier les cellules vides tel que NB. VIDE ou les marqueurs de valeurs manquantes (ex.: N/A, NULL). Il est aussi possible d'obtenir ces informations en filtrant les colonnes.
- ⊕ **Valeurs aberrantes** : Détecter les valeurs qui s'écartent significativement des autres tel que des valeurs numériques décuplées.
- ⊕ **Incohérences et erreurs de saisie** : Rechercher des incohérences logiques (dates impossibles, textes mal formatés) et des erreurs de saisie courantes.

Une fois identifiées, ces données problématiques doivent être traitées pour améliorer la qualité et la fiabilité des analyses subséquentes.

4. Nettoyage des Données

Le nettoyage des données est une étape cruciale pour garantir la qualité, l'exactitude et la fiabilité des analyses ultérieures. Cette étape consiste principalement à identifier et traiter les données manquantes, les valeurs aberrantes, et les incohérences pour préparer un jeu de données propre et utilisable.

Identifier les données manquantes

Les données manquantes sont un problème fréquent dans tout ensemble de données et peuvent survenir pour diverses raisons, telles que des erreurs de saisie ou des problèmes de collecte. Voici comment les traiter :

- Utiliser des techniques de comptage et des visualisations pour repérer les cellules vides ou marquées comme NA, NULL, ou tout autre indicateur de valeurs manquantes.
- Effectuer un audit des colonnes pour déterminer la proportion de données manquantes par colonne ou par ligne.

Déterminer si les données manquantes sont obligatoires ou pertinentes

- Évaluer l'importance des données manquantes pour votre analyse. Si les données manquantes ne sont pas cruciales, il peut être possible de les ignorer sans impact significatif sur les résultats.
- Analyser la pertinence des données manquantes en fonction des objectifs de l'analyse.

Traitement des données manquantes

- **Suppression des valeurs manquantes** : Si les données manquantes représentent une faible proportion du jeu de données et sont réparties de manière aléatoire, les lignes ou colonnes contenant des valeurs manquantes peuvent être supprimées sans compromettre l'intégrité de l'ensemble de données.
- **Imputation** : Lorsque les données manquantes sont importantes, des techniques d'imputation peuvent être utilisées :

    - ⊕ Moyenne, médiane ou mode : Remplacer les valeurs manquantes par la moyenne, la médiane ou le mode des valeurs de la colonne.
    - ⊕ Valeurs précédentes ou suivantes : Utiliser les valeurs des observations précédentes ou suivantes pour combler les lacunes (méthode de remplissage forward/backward).

Archiver les lignes avec des données manquantes

Si l'option retenue est de ne pas traiter les données manquantes mais de les archiver, voici comment procéder :

1. **Identifier les lignes à archiver**

- Filtrer les lignes qui contiennent des valeurs manquantes dans les colonnes critiques ou au-delà d'un seuil acceptable de valeurs manquantes.

2. **Méthodologie d'archivage**

- Créer un fichier d'archive : Séparer les lignes contenant des valeurs manquantes et les sauvegarder dans un fichier distinct pour une éventuelle utilisation future ou pour la documentation.
- Documenter le processus : Noter les raisons pour lesquelles ces lignes ont été archivées et les critères utilisés pour cette décision. Inclure des métadonnées telles que la date d'archivage, les colonnes affectées, et toute autre information pertinente.
- Mettre à jour le jeu de données principal : Supprimer les lignes archivées du jeu de données principal et vérifier que cette opération n'a pas introduit de nouvelles incohérences ou erreurs.

3. **Conserver les archives de manière sécurisée**

- Assurer la sécurité et l'intégrité des fichiers d'archive en utilisant des méthodes de stockage appropriées, comme le chiffrement et les contrôles d'accès restreints.

En suivant ces méthodologies, vous pouvez garantir que les données manquantes sont traitées de manière appropriée, améliorant ainsi la qualité et la fiabilité de votre jeu de données pour des analyses futures.