Rapport de traitement des données

Titre du projet	Préconisation, étapes et traitement des données afin de garantir le respect du RGPD Dev' Immédiat.

Version	Auteur	Description	Date
V1	*François*	*Rapport*	*01.06.2022*

Introduction

Dev' Immédiat, courtier en assurance automobile, a été sanctionné par la CNIL suite à une plainte client sur le non-respect des règles RGPD concernant la gestion et l'utilisation des données personnelles de ses clients.
En effet le non- respect des règles notamment sur la collecte et la conservation de très vielles données personnelles obsolètes a eu pour effet de traiter une demande devis commercial de manière complétement erronée et biaisée.

L'Equipe Dev' Immédiat a été complétement paniqué et pris au dépourvu sur la manière de comment traiter la demande d'accès aux données personnelles étant donné le manque totale de process pour gérer le cycle de vie des données et l'absence de politique de gestion de la protection des données de bout en bout chez Dev Immédiat.

Le besoin est donc de mettre en place immédiatement des actions correctives et une politique claire de gestion des données personnel pour être conforme aux règles du RGPD afin d'être en règles vis-à-vis de la CNIL avec l'enjeux de pouvoir remettre au plus vite à disposition de l'équipe de performance commerciale une nouvelle extraction des données du CRM en toute légalité, conforme aux règles RGPD et ainsi pouvoir lever la sanction.

L'objectif de l'entreprise est donc de revoir les processus internes pour la collecte et le traitement des données clients, la documentation, la communication client, les processus et l'accès aux données personnelles afin d'être conforme et en règle vis-à-vis de la CNIL.
Le Bénéficie à terme étant bien entendu de permettre de garantir à l'avenir le respect du RGPD Dev' Immédiat.

Préconisations en lien avec les règles RGPD permettant de garantir à l'avenir le respect du RGPD

- **Mettre en place une communication claire et transparente avec le Client sur la collecte l'utilisation et le traitement de ses données personnelles ainsi que leurs droits afin qu'il donnent explicitement leur consentement :**

    - Mise en place d'un processus d'information (politiques de confidentialité) et de consentement clair (formulaires) et transparent lors de la saisie des informations du client sur le site web et le dépôt de cookies
    - Expliquer sous forme d'accès pop-up quelles données sont collectées, l'équipe Dev'Immediat doit réfléchir à la finalité du pourquoi ces données sont collectées (est-ce par ex. pour établir le meilleurs tarifs d'assurance auto au plus près des usages quotidien ?) et comment elles sont utilisées (à des fins commerciales statistiques, suivi de dossier etc..).
    - Il faut également informer les clients des droits dont ils disposent en matière de protection des données, tels que le droit d'accès, de rectification, d'effacement et d'opposition sur les données stockées dans le CRM Dev'Immediat

- **Mettre la base de données du CRM conforme au RGPD au niveau de la collecte :**

    - Une étude détaillé doit être menée pour ne sélectionner que les données strictement nécessaires et collectées pour le domaine de l'assurance auto (comme le type et l'usage du véhicule), des infos obsolètes des années 80 comme la couleur rouge de la voiture doivent être supprimés.
    - Les données sensibles qui ne peuvent être justifiées par les bases légales fixées par le RGPD Dev Immédiat ne doivent pas être collectées et être supprimées du CRM (par ex. numéro de Sécurité Sociale, le Groupe sanguin, le nom de l'employeur, …).
    - Des catégories de données, surement nécessaire à des fins statistiques, sont trop intrusives et doivent être revues. Par exemple, le revenu doit être remplacé par des tranche de revenus et le métier pourrait être remplacé par des catégories socio-professionnelles.

- **Etablir un traitement sur les données personnelles conformes au RGPD afin de faciliter le partage des données entre service nécessaires aux différentes analyses internes :**

    - Donnée d'identification direct à supprimer: nom, prénom, email…
    - Un traitement d'anonymisation doit être mis en œuvre pour protéger la confidentialité des individus et éviter la divulgation d'informations sensibles ou confidentielles tout en permettant l'exploration et la prise de décision basée sur les données.
    - Ainsi des tranches de valeurs plutôt que des valeurs explicites doivent être mise en place pour masquer et anonymiser les données : la date de naissance précise des clients doit être remplacée par des catégories générales : Jeune : moins de 25 ans /

Adulte : 25-64 ans / Senior : 65 ans et plus), le nb d'enfant en conduite accompagné, le nombre de points du permis perdus…

Analyse des attributs des données CRM pour être en accord avec le RGPD

Nom Attribut Collecté	Analyse effectuée pour la véracité de la donnée en lien avec le RGPD
metier	• A supprimer en amont • Considéré comme trop intrusif et peut être rebouclé avec la ville pour identifier directement une personne • A l'avenir doit être remplacé par des catégories socio-professionnelles
employeur	• A supprimer en amont • Donnée sensible non lié à l'assurance Auto
Num_ss	• A supprimer en amont • Donnée sensible médical.
Groupe_sanguin	• A supprimer en amont • Donnée sensible médical.
Id_site_web	• A supprimer en amont • Le même identifiant permet de recroiser une même personne sur plusieurs type de données.
Nom	• A supprimer en amont • Donnée personnel identifiant directement la personne
sexe	• A garder pour statistique éventuelle en fonction des besoins, difficilement recroisable vu le nombre d'échantillon.
email	• A supprimer en amont • Donnée personnel identifiant directement la personne
date_naissance	• A garder mais à traiter et remplacer pour anonymiser sous forme de tranche (Jeune : moins de 25 ans / Adulte : 25-64 ans / Senior : 65 ans et plus)
id_client	• A supprimer en amont • Identifiant permettant de reboucler directement avec la personne • Générer un id aléatoire à la place
enfant_conduite_accompagne	• A garder mais à traiter et remplacer pour anonymiser sous forme de tranche un", "plusieurs" ou "aucun"
nombre_enfants	• A garder mais à traiter pour rendre l'info binaire (oui /non)
revenus	• A garder pour statistique éventuelle mais à traiter et remplacer sous forme de tranche [x -yK [
valeur_residence_prin	• A supprimer en amont • Donnée sensible non liée à l'assurance Auto
formation	• A garder pour statistique éventuelle en fonction des besoins (+ nettoyer mauvaise typo)
usage_vehicule	• A garder • Donnée liée à l'assurance Auto, non sensible
type_vehicule	• A garder (+ nettoyer mauvaise typo) • Donnée liée à l'assurance Auto, non sensible
est_rouge	• A supprimer en amont • Préjugé et faux cliché des années 80 de la couleur impactant les assurances auto.
points_perdus	• A garder mais à traiter et remplacer pour anonymiser sous forme de tranche faible", "moyen" ou "élevé"

	• Donnée liée à l'assurance Auto
age_vehicule	• A garder • Donnée liée à l'assurance Auto, non sensible
type_conduite	• A garder (+ nettoyer mauvaise typo) , non sensible • Donnée liée à l'assurance Auto, non sensible
date_demande	• A garder pour le traitement et le calcul de l'âge de l'assuré (et automatisation futur du traitement), pourrait être supprimé par la suite.
etat_dossier	• A supprimer en amont • Uniquement les état complet sont traités
formule	• A garder • Donnée liée à l'assurance Auto
tarif_devis	• A garder mais à traiter et remplacer pour masqué les détails tarifaire sous forme de Tranche Tarifaire: bas, moyen et élevé
adresse	• A garder mais à traiter et remplacer pour anonymiser pour ne garder que le code postal d'habitation (tarif assurance dépend du lieu ville ou campagne) car on ne peut anonymiser plus en l'état (plutôt un recodage + poussé avec taille de la ville)
lat	• A supprimer en amont • Donnée personnel identifiant directement la localisation de la personne
lon	• A supprimer en amont • Donnée personnel identifiant directement la localisation de la personne

Le but étant d'adapter le jeu de données de façon à ce qu'on ne puisse pas croiser les informations pour remonter à l'utilisateur :

Les attributs ne respectant donc les règles de RGPD peuvent entre exclus directement en amont à partir de la requête SQL de récupération des données du CRM.

Les attributs restants sont donc nécessaires au traitement automobiles, soit directement ou temporairement (préparation) afin de les retravailler pour l'étape d'anonymisation

Récupération des données du CRM par requête SQL

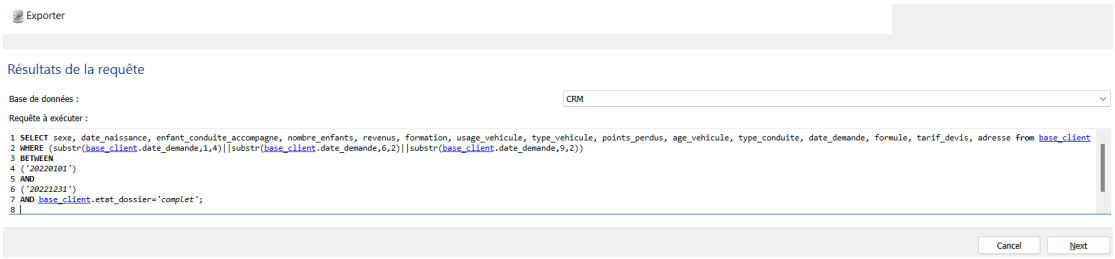
Extraction des données de l'année **2022** pour les clients dont l'état du dossier est « **complet** » à partir de la base **SQL.**

```
SELECT sexe, date_naissance, enfant_conduite_accompagne, nombre_enfants, revenus, formation, usage_vehicule, type_vehicule, points_perdus, age_vehicule, type_conduite, date_demande, formule, tarif_devis, adresse from base_client
WHERE
(substr(base_client.date_demande,1,4)||substr(base_client.date_demande,6,2)||substr(base_client.date_demande,9,2))
BETWEEN
('20220101')
AND
('20221231')
AND base_client.etat_dossier='complet';
```

Clause du Select : je ne sélectionne que les attributs qui fond du sens pour le fichier commercial de l'assurance auto ou ceux nécessaires au traitement avec Power Query pour la préparation et l'anonymisation les données. (attribut du tableau précédent)

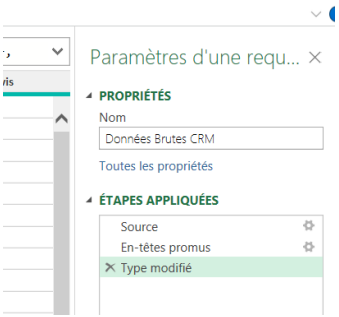
Clause du Where:
je filtre sur l'extraction de l'attribut "date_demande" en ne prenant que les 4 premier caractères (position 1 de la chaine) correspondant à l'année, concaténé au 2 caractères suivant (position 6 de la chaine) correspondant au mois, concaténé au 2 caractères suivant (position 9 de la chaine) correspondant au jour. Cette extraction doit être compris entre 20220101 et 20221231 pour ne conserver que les records de l'année 2022.
Et je ne prends ensuite que les dossier à l'état complet.

La requête SQL est donc directement exécuter dans SQLite pour l'extraction des données brutes CRM souhaitées au format CSV.



Etapes des traitements effectués pour la préparation et l'anonymisation du jeu de données avec l'ETL Power Query :

Chargement des Données Brut CRM provenant de l'extraction de la requête SQL.

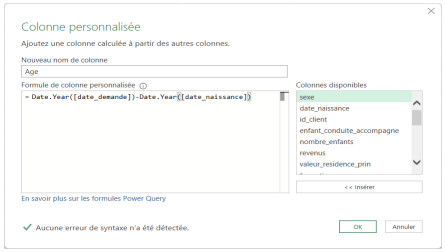


Les en-tête sont automatiquement reconnues et promues à partir de la 1 er ligne du fichier CSV.

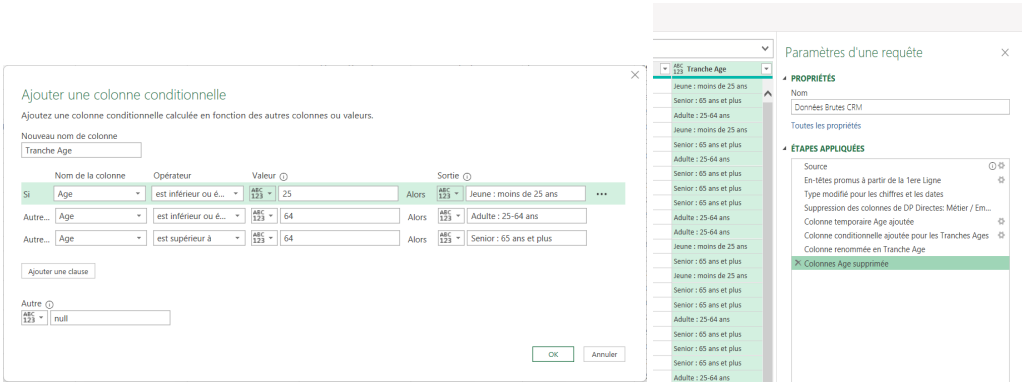
Les type sont automatiquement modifiée pour les données nombre entier uniquement (non decimal sans « . » et les dates (ou date horaire).

Transformation l'information très personnelle et potentiellement « croisable » de la date de naissance en tranche d'âge

- Calcul de l'âge de la personne à partir de la date de la demande du devis

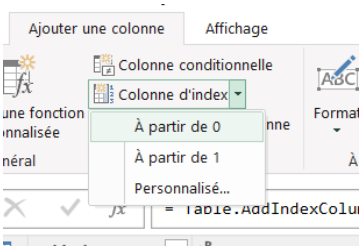


- Split de la nouvelle colonne Age en nouvelle colonne conditionnelle par tranches d'âges (Jeune : moins de 25 ans / Adulte : 25-64 ans / Senior : 65 ans et plus) et remplacement de la date de naissance



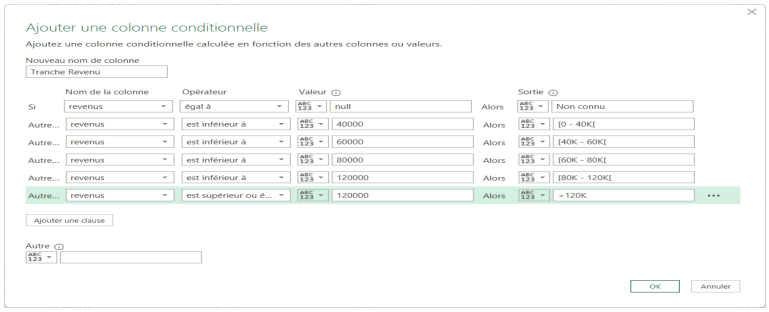
Ajout colonne d'index unique incrémental

- Le nouvel index aléatoire généré évite de pouvoir recroiser avec d'autre base le même identifiant (ce qui n'était potentiellement pas le cas de id_client précédemment).

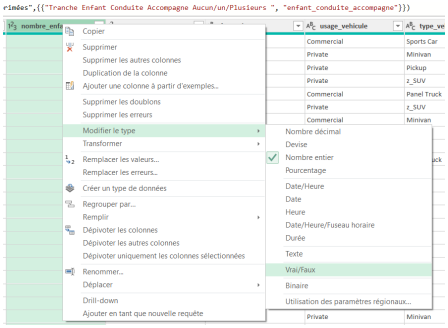


Anonymiser des données très détaillés : nb d'enfant en conduite accompagné, les Revenus et le nombre de points perdu en les splittant plutôt par Tranche plus globales et plus anonymes

- Remplacement avec la nouvelle colonne enfant_conduite_accompagne conditionnelle par tranches un", "plusieurs" ou "aucun".
- Remplacement du revenu avec la nouvelle colonne conditionnelle « Tranche Revenu » par tranches : Non connu / [0 - 40K[/ [40K - 60K[/ [60K - 80K[/ [80K - 120K[/ >120K
- Remplacement du points_perdus avec une nouvelle colonne conditionnelle par tranches : faible", "moyen" ou "élevé"

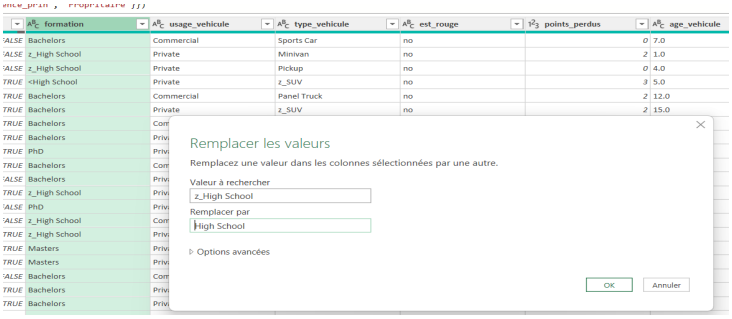


Masquer et remplacer l'information très détaillé du nombre d'enfant en information générale booléenne



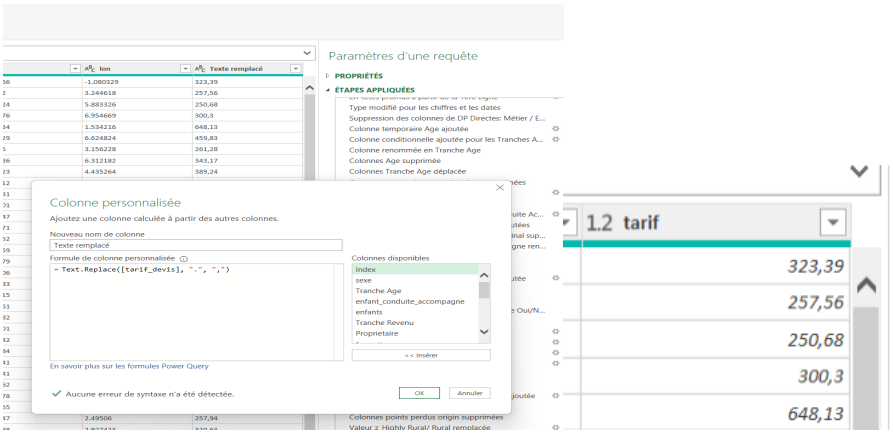
Nettoyer les erreurs de typo ou mauvaises saisies.

- Nettoyage des préfixes « z » ou « < » devant les mauvaise saisies pour le formation type_vehicule et type_conduite



Préparation du jeu de données pour empêcher que l'équipe de performance commerciale n'ait accès aux données tarifaires détaillées

- Création d'une nouvelle colonne temporaire Tarif afin de remplace le point « . » du tarif_devis de format texte en virgule « , » afin de pouvoir ensuite convertir le format en décimal
- Transformation du type de la colonne temporaire Tarif en décimal

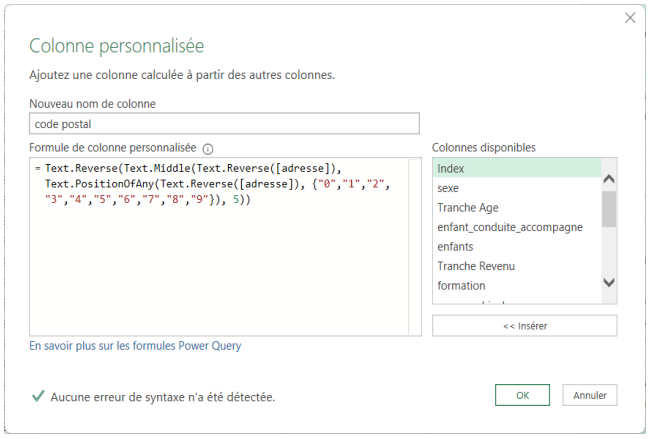


- Création d'une nouvelle colonne conditionnel Tranche Tarifaire afin de regrouper les différents tarifs par tranche : « Basse » ou « Moyenne » ou « Elevée »

- Suppression de l'ancienne colonne tarif_devis
- Suppression de la colonne temporaire Tarif

Anonymisation de l'adresse complète pour ne garder que le code postal d'habitation

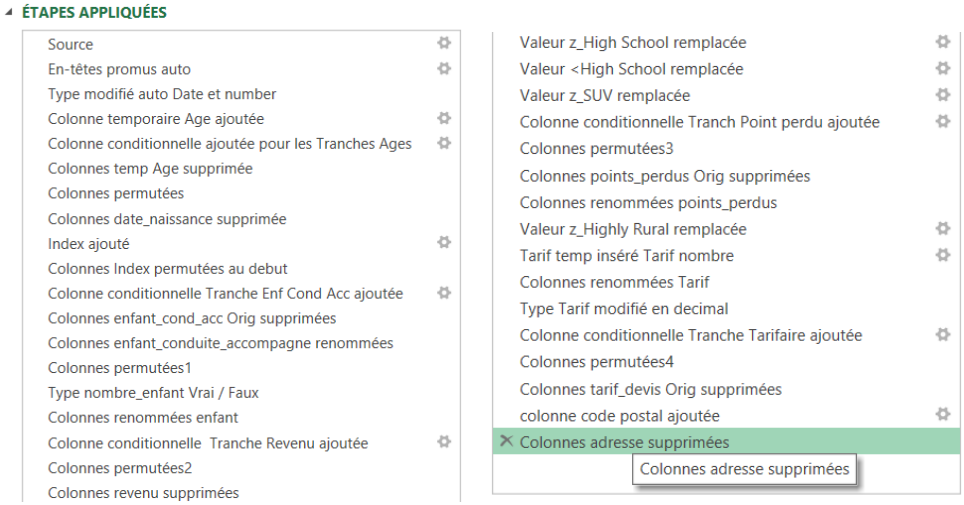
- Le lieu d'habitation (urbain ou rurale), étant une information très importante, le code postal permet de masquer l'adresse complète personnel du client



- L'extraction du code postal s'effectue en inversant la chaine de caractère et en récupérant la position du premier chiffre du code postal pour en extraire ensuite les 5 chiffres que compose le code postal.

Resumé des étapes effectuées dans Power Query :

- Le code M correspondant à toutes les étapes et les traitement effectués (requête PQ) peut être sauvegardé et réappliqué automatiquement sur un nouveau jeux de données(par exemple sur un extrait CSV de l'année 2023)



Préconisations à faire pour les principes qui ne seraient pas immédiatement applicables sur les données.

- Au niveau de la collecte des données clients, via le formulaire de saisie , il faudrait remplacer le métier par des catégories socio-professionnelles beaucoup plus générique afin d'éviter de pouvoir recroiser et d'identifier (un seul corps de métier dans une ville).

- De plus le code postal devrait être retravaillé et retraité avec le nombre moyen d'habitants pour plutôt fournir une information beaucoup plus générale de « grandeur-taille » du lieu d'habitation nécessaire aux analyses de cout de devis d'assurance (Grande Ville, Moyenne, Petite, Bourg, Campagne etc…) afin d'éviter de recroiser des informations comme avec le sexe par exemple (1 femme habitant à ce code postal de tel tranche d'âge pourrait éventuellement être reconnu).

- Réfléchir mettre à jour et revoir le jeux de données à collecter complémentaire, plus pertinent lié à assurance automobile avec les équipes commerciales : par exemple la Tranche de kilomètre estimé parcouru par an, la Motorisation du véhicule: Electrique / hybride / essence / gasoil, la puissance en de la voiture, etc…

- Procéder plus tard à une évaluation de l'impact sur la protection des données (EIPD) pour identifier les risques potentiels liés au traitement des données personnelles. Cela aidera à déterminer les mesures de sécurité supplémentaires à mettre en place pour garantir la protection des données.

- Former le personnel Dev' Immédiat : Organiser des séances de formation pour sensibiliser les employés aux principes du RGPD, aux exigences de confidentialité et aux pratiques recommandées en matière de traitement des données personnelles.

- Effectuer des audits de conformité : Réaliser régulièrement des audits internes Dev' Immédiat pour vérifier la conformité aux principes du RGPD. Ces audits doivent inclure des évaluations des procédures, des mesures de sécurité et des systèmes de gestion des données pour s'assurer qu'ils respectent les exigences du RGPD afin d'établir des protocoles d'urgence pour faire face aux violations de données et aux incidents de sécurité.

Exigences, contraintes et enjeux de la confidentialité des données

- La confidentialité des données est donc régies par des exigences légales et des réglementations juridique permettant d'encadrer le traitement des données personnelles sur tout le territoire de l'Union européenne : le Règlement Général sur la Protection des Données Cela inclut le consentement éclairé des individus concernant l'utilisation de leurs données personnelles, le droit à l'effacement des données, la notification de violations de données, etc.

    ⇨ L'explication de ces exigences permet de sensibiliser les individus à la nécessité de protéger leurs données personnelles.

- Ainsi la confidentialité des données met en lumière l'importance de préserver la confiance des utilisateurs dans le traitement de leurs données personnelles. Le fait d'expliquer clairement comment la confidentialité des données contribue à la confiance des individus dans une organisation peut aider à établir des relations de confiance durables avec les clients et les partenaires.

- La confidentialité des données impose aussi des contraintes opérationnelles fortes et met l'accent sur des aspects organisationnels et techniques à prendre en compte pour garantir la confidentialité des données : de la mise en place de mesures de sécurité appropriées pour protéger les données à la formation du personnel sur les bonnes pratiques en matière de protection des données ou encore l'utilisation d'outils de chiffrement pour sécuriser les données sensibles.

- Un des enjeux de la confidentialité des données est le potentiel risque de violations de données : cela peut inclure des risques tels que la perte de données, la fuite d'informations sensibles (piratage). Si ces données tombent entre de mauvaises mains ou sont utilisées de manière abusive, cela peut entraîner des conséquences néfastes pour les individus concernés, comme la fraude, l'usurpation d'identité, ou la violation de leur vie privée Les conséquences juridiques et financières qui en découlent peuvent être dramatiques. L'évaluation attentive de ces risques permet d'anticiper et de mettre en place des mesures de protection adéquates afin de limiter au mieux la violation de la confidentialité des données.