De l'ETL aux Analyses de Marché et Conformité RGPD

I. Quels sont les principes de la RGPD ?

Le règlement général sur la protection des données (RGPD) fixe plusieurs principes de base sur le traitement et la publication des données personnelles.

1. Licéité, limitation des finalités, minimisation des données

- **Licéité :** les données personnelles ne doivent être collectées que pour des finalités légitimes, explicites et déterminées.
- **Limitation des finalités** : les données collectées ne peuvent être utilisées que les finalités pour lesquelles elles ont été collectées.
- M**inimisation des données** : la quantité de données doit être réduite au minimum nécessaire pour atteindre les finalités déterminées.

En d'autres termes, les entreprises doivent être transparentes sur les raisons pour lesquelles elles utilisent les données. Elles ne doivent stocker et utiliser que les données nécessaires à leur besoin. Les mentions légales permettent d'informer les utilisateurs de la collecte et la finalité de l'utilisation de ses données.

2. Exactitude

Les données personnelles doivent être exactes et, si nécessaire, mises à jour. Des mesures doivent être prises pour que les données inexactes soient rectifiées ou effacées. Cela signifie que les entreprises doivent prendre les mesures nécessaires pour s'assurer que les données personnelles qu'elles détiennent sont à jour.

3. Principe de limitation du traitement

Les données personnelles ne doivent être conservées que pendant une durée proportionnelle aux finalités pour lesquelles elles sont traitées. En d'autres termes, les entreprises doivent avoir une politique claire de conservation des données et supprimer les données lorsqu'elles ne sont pas nécessaires.

4. Intégrité et confidentialité

Les données personnelles doivent être traitées de façon à garantir leur sécurité, y compris la protection contre le traitement non autorisé ou illicite et contre le perte, la destruction ou le dommage accidentels, par des mesures techniques ou organisationnelles adéquates.

Cela signifie que les entreprises doivent mettre en place des mesures de sécurités adaptées pour protéger les données personnelles contre les accès non autorisés, les utilisations abusives et les pertes accidentelles.

5. Responsabilités

Le responsable du traitement est responsable du respect des principes susmentionnés et doit être en mesure de démontrer sa conformité. Les entreprises doivent désigner un responsable de la protection des données (DPO) qui sera responsable de la mise en œuvre et du respect du RGPD.

6. Respect des droits des personnes

Les personnes ont le droit d'accéder à leurs données personnelles, de les rectifier, de les supprimer, de les effacer, de limiter leur traitement, de s'opposer à leur traitement et de les transférer. Les entreprises doivent informer leurs clients de leurs droits en matière de protection des données et leur permettre d'exercer ces droits facilement.

7. Protection des données dès la conception et par défaut

Le responsable du traitement des données doit mettre en place des mesures techniques et organisationnelles appropriées pour protéger les données personnelles dès la conception du traitement et tout au long des produits et services. Les entreprises doivent intégrer le traitement des données dans leur stratégie globale d'entreprise.

II. Les différentes techniques de protection des données personnelles

Pour respecter les exigences RGPD, il existe plusieurs techniques de protection des données avant leur diffusion en Open Data :

- **Anonymisation :** L'anonymisation vise à supprimer tout lien entre les données publiées et les personnes concernées, rendant ainsi impossible l'identification directe ou indirecte. Les méthodes incluent :
    - Généralisation : Agréger les données pour masquer des détails spécifiques (ex. : âge par tranche d'âge plutôt que l'âge exact).
    - Randomisation : Ajouter des perturbations ou des « bruits » aux données pour brouiller les informations sans altérer les analyses statistiques.
- **Pseudonymisation :** Cette technique consiste à remplacer les informations identifiantes par un identifiant fictif. Contrairement à l'anonymisation, la pseudonymisation n'empêche pas la ré-identification, mais elle en réduit le risque et le facilite pour des utilisateurs autorisés via une clé de déchiffrement.
- **Masquage des données** : Technique souvent utilisée pour masquer partiellement des informations sensibles, par exemple les numéros de téléphone ou d'autres identifiants.

Ces méthodes doivent être choisies en fonction de la sensibilité des données et de l'objectif de la publication, en assurant un compromis entre sécurité et utilité des données.

Sécurité des Données pour Prévenir les Fuites d'Informations Sensibles

La sécurité des données en Open Data implique la mise en place de mesures de prévention contre les accès non autorisés et les fuites d'informations :

- **Contrôle des accès et authentification** : Utiliser des systèmes d'authentification robustes pour les données non ouvertes, comme des mots de passe forts, une authentification à deux facteurs, et restreindre les accès aux données brutes avant publication.
- **Chiffrement des données** : Les données sensibles, même dans les systèmes de stockage temporaire, doivent être chiffrées pour éviter tout risque d'exposition en cas de fuite.
- **Suivi et audit** : Mettre en place des mécanismes de journalisation et d'audit pour suivre l'accès et les modifications apportées aux données, garantissant ainsi une traçabilité des opérations.
- **Tests de vulnérabilité** : Effectuer régulièrement des tests de sécurité pour identifier et corriger les failles potentielles dans les systèmes de stockage et de diffusion des données.
- **Formation des équipes** : Sensibiliser les équipes aux meilleures pratiques de sécurité est crucial pour assurer le respect des règles de sécurité dans le traitement et la publication des données.

Secret Statistique pour Protéger les Données Agrégées

Le secret statistique est une exigence légale visant à protéger les données agrégées contre la ré-identification des individus dans les statistiques publiques ou les bases de données anonymisées. Les méthodes principales incluent :

- **Seuil de diffusion** : Fixer un nombre minimal d'observations pour publier des données agrégées afin de garantir que les individus ne puissent être identifiés.
- **Suppression de certaines données** : Supprimer les sous-catégories présentant un risque de ré-identification, notamment dans les groupes de petite taille.
- **Perturbation contrôlée** : Ajouter des erreurs contrôlées aux statistiques pour empêcher l'identification des données sensibles tout en conservant l'utilité statistique des données publiées.
- **Contrôle de la granularité** : Limiter la précision des données, par exemple en publiant des informations sur des régions plutôt que sur des municipalités lorsque celles-ci sont très petites.

II. Qu'est ce qu'une base OpenData ?

L'Open Data, ou données ouvertes, désigne la mise à disposition libre et gratuite de données publiques ou privées, accessibles pour tous, afin d'être consultées, réutilisées et partagées sans restrictions. Cette pratique repose sur l'idée que la transparence et le partage d'informations peuvent favoriser l'innovation, la participation citoyenne, et le développement économique. L'Open Data permet ainsi aux chercheurs, entreprises, développeurs et citoyens de créer de nouveaux services, d'analyser les données pour prendre de meilleures décisions, et de rendre les gouvernements plus responsables. Par exemple, dans le domaine des transports, les données ouvertes sur les horaires et trajets permettent à des applications comme Google Maps de proposer des itinéraires en temps réel. Bien que prometteuse, l'Open Data pose aussi des défis, notamment en matière de protection de la vie privée et de qualité des données, nécessitant des normes claires et une gouvernance efficace pour garantir un accès équitable et éthique.Il est donc important que l’Open Data respecte les règles imposées par la RGPD.

III. Base de données relationnelles, analytiques et Open Source.

1. Bases de données relationnelles (Relationnelles)

Les bases de données relationnelles sont les plus couramment utilisées pour le stockage de données structurées et organisées. Elles se basent sur le modèle relationnel, où les données sont stockées dans des tables interconnectées par des relations. Le modèle suit les principes de normalisation, qui limitent la redondance des données, assurent leur cohérence et facilitent les mises à jour. Le langage de requête SQL (Structured Query Language) est utilisé pour manipuler et interroger ces bases.

- **Cas d'utilisation** : Idéales pour les applications transactionnelles, comme les systèmes de gestion d'inventaire, les logiciels de comptabilité, ou les applications bancaires, où des opérations fréquentes de lecture et d'écriture sont nécessaires.
- **Avantages** : Cohérence des données, intégrité référentielle, flexibilité des requêtes.
- **Inconvénients** : Les bases relationnelles peuvent devenir peu performantes pour des volumes de données très importants et pour des analyses complexes, car les jointures de tables consomment des ressources et peuvent ralentir les requêtes.

2. Bases de données analytiques

Les bases de données analytiques, souvent appelées entrepôts de données (data warehouses), sont spécialement conçues pour les requêtes de type analytique et le traitement des gros volumes de données. Elles sont optimisées pour les lectures intensives et la génération de rapports, plutôt que pour les transactions fréquentes. Ces bases de données sont souvent dénormalisées, ce qui permet de limiter les jointures complexes, et utilisent des structures en étoile ou en flocon pour organiser les données. Le but est de faciliter l’analyse de données multidimensionnelles pour obtenir des insights rapides et pour répondre aux besoins décisionnels.

- **Cas d'utilisation** : Très utilisées dans la business intelligence, le reporting, et les analyses de données historiques pour des entreprises qui doivent tirer des enseignements de données sur le long terme.
- **Avantages** : Excellentes performances pour les requêtes analytiques, organisation facilitée pour les agrégations et les calculs avancés.
- **Inconvénients** : Moins adaptées pour les transactions fréquentes en temps réel, coûts de maintenance et de stockage élevés pour les gros volumes de données.

3. Bases de données agrégées de type Open Data

Les bases de données agrégées, utilisées pour les jeux de données Open Data, se concentrent sur la publication et l’accessibilité de données en tant qu’ensemble de valeurs brutes (datasets) ou agrégées. Ces bases sont généralement dénormalisées et structurées de manière à être facilement exportées, partagées et consommées par des tiers. Les données sont souvent fournies sous des formats standardisés comme CSV, JSON ou XML, facilitant leur utilisation par les développeurs et les analystes de données dans leurs propres outils ou applications.

- **Cas d'utilisation** : Idéales pour des données publiques et ouvertes, comme les statistiques gouvernementales, les données de transport public, ou les informations environnementales, où l'objectif est de faciliter la transparence et l'utilisation des données.
- **Avantages** : Facilité d'accès, compatibilité avec de nombreux formats et outils, favorise la réutilisation des données pour la recherche, le développement d'applications et l'analyse.
- **Inconvénients** : Limitées en termes de cohérence et de contrôle de la qualité des données, potentiels problèmes de mise à jour en temps réel, et souvent peu adaptées aux requêtes complexes nécessitant des relations entre les données.

Aspect	Relationnelles	Analytiques	Open Data
Type de données	Structurées et normalisées	Structurées, souvent dénormalisées	Agrégées ou brutes, standardisées
Cas d'utilisation	Applications transactionnelles	Business intelligence, analyses historiques	Publication de données publiques
Performance	Optimisées pour les transactions rapides	Optimisées pour les requêtes analytiques	Accès rapide aux de données statistiques
Structure	Tables interconnectées	Structure en étoile ou flocon	Fichiers plats (XLM, CSV, XLSX…)
Langages de requête	SQL	SQL	Accès direct
Avantages	Cohérence, intégrité, flexibilité des requêtes	Optimisées pour les analyses multidimensionnelles	Accessibilité, compatibilité, réutilisation
Inconvénients	Peu performantes pour gros volumes analytiques	Pas conçues pour les transactions fréquentes	Contrôle de qualité et mise à jour complexes

IV. Index et partition de table

1. Importance des Index

Un index est une structure de données qui améliore la vitesse des opérations de recherche sur une table en permettant un accès rapide aux lignes en fonction de valeurs spécifiques dans une colonne (ou un ensemble de colonnes). En analogie, un index est comme la table des matières d'un livre : il permet d'accéder directement à l'information sans devoir parcourir chaque page.

- **Amélioration des performances** : Les index réduisent le nombre de lectures de disque nécessaires pour retrouver des données, surtout pour des requêtes de recherche fréquentes ou pour celles utilisant des filtres.
- **Optimisation des jointures :** Dans des bases de données relationnelles où les jointures entre tables sont fréquentes, les index sur les colonnes de clés étrangères et de clés primaires peuvent grandement réduire le temps de traitement, car ils permettent d'accéder rapidement aux lignes correspondantes dans les tables reliées.
- **Inconvénients des index :** Les index ne sont pas sans coût. Ils augmentent la taille de la base de données et nécessitent des mises à jour lorsque des opérations d’insertion, de suppression ou de mise à jour sont effectuées sur les données. Trop d’index peuvent également ralentir les opérations d’écriture.