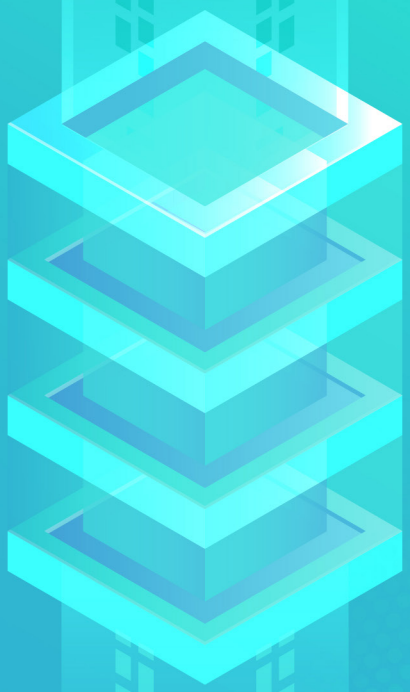
LIVRE BLANC

QUALITÉ DES DONNÉES

UN LEVIER STRATÉGIQUE POUR VOTRE ENTREPRISE



alliance digitale :
dma France
MMAf
iab france

Sommaire

Préambule

La qualité des données, un enjeu capital à l'ère du big data et de l'IA

Les données sont le véritable carburant de l'économie numérique, un capital devenu essentiel que les organisations doivent valoriser. Chaque jour, des millions de données sont générées et utilisées *via* une multitude de canaux présents sur les **parcours transactionnels,** les **interactions client**s ou encore les **plateformes web et mobiles**. L'intelligence artificielle, grande consommatrice de données pour ses apprentissages, en génère également de nouvelles (textes, images, prédictions). Cette abondance entraîne une grande hétérogénéité des données, avec des formats, des structures et des niveaux de fiabilité très variables.

LES COÛTS CACHÉS ET L'IMPACT FINANCIER DE LA NON-QUALITÉ

Des données incomplètes, incohérentes ou obsolètes représentent un **risque bien plus qu'un atout** pour l'entreprise. Une qualité insuffisante de données peut déclencher l'effet papillon, en **impactant les processus opérationnels, les décisions, la confiance des clients** et même **en biaisant les résultats** des projets d'IA. Avec des conséquences directes néfastes pour la gestion client, les résultats des campagnes marketing ou le fonctionnement des outils d'IA. Il est aisé de déduire que l'impact financier de la non-qualité peut vite devenir considérable.

OBJECTIFS ET PÉRIMÈTRE DE CE LIVRE BLANC

C'est pour vous aider à bâtir une démarche de qualité des données que ce livre blanc a été conçu.

Pour vous accompagner, pas à pas, ce document :
- **pose les concepts clés** de la qualité des données ;
- **présente de bonnes pratiques**, illustrées par des cas concrets ;
- **montre comment mesurer, corriger et piloter** efficacement la qualité des données ;
- **explique comment réduire les coûts,** améliorer les décisions et sécuriser les projets d'IA.

Ce document s'adresse aux spécialistes de la donnée, mais également aux responsables de la gouvernance des données, aux équipes IA et à tous les décideurs. De même, toute personne ayant un intérêt pour la valorisation des données trouvera dans ce livre blanc une utilité.

Partie 1.

Les fondamentaux de la qualité des données

1. Qu'est-ce que la qualité des données ?

La **« qualité des données »** (*data quality* en anglais) désigne **le degré selon lequel un ensemble de données répond aux besoins de ses utilisateurs.** On l'évalue traditionnellement selon **neuf critères « intrinsèques et essentiels » :**

- **La complétude :** toutes les informations attendues sont présentes.
- **L'unicité :** absence de doublons.
- **L'exactitude :** correspondance avec la réalité.
- **La validité :** conformité aux formats et règles définis.
- **La cohérence :** absence de contradiction entre les sources.
- **La fiabilité :** concerne la qualité de la source et du traitement.
- **L'actualité :** les données sont assez récentes.
- **La précision :** le niveau de détail de la donnée est suffisant pour répondre au besoin.
- **L'uniformité :** homogénéité des formats, des unités et des conventions utilisées.

Ces critères sont complétés par **huit critères « contextuels et contributifs » :**

- **La traçabilité :** historique clair des transformations.
- **La clarté :** données compréhensibles et bien documentées.
- **L'utilité :** pertinence vis-à-vis du besoin réel.
- **L'accessibilité :** facilité d'accès pour les utilisateurs autorisés.
- **La conformité :** respect des réglementations.
- **La sécurité :** protection contre les accès et usages non autorisés.
- **La disponibilité :** données accessibles au bon moment.
- **La pérennité :** capacité à conserver la donnée dans le temps.

À chaque nouvelle collecte de données, il est possible d'évaluer chacun de ces critères et de s'orienter vers une qualité maximale en se posant les questions suivantes : **mes données seront-elles complètes et uniques ? Mon nouvel ensemble de données est-il cohérent et fiable ? D'où viennent-elles ? Sont-elles conformes aux règlementations ? Sont-elles en sécurité et facilement accessibles et disponibles pour les utilisateurs ?**

Dans certains cas, dès la collecte, des méthodes et des outils permettent d'optimiser la qualité des données. Dans d'autres cas, il sera préférable de mettre en œuvre des processus dans le système d'information. Parfois, il ne faut pas hésiter à rejeter ou exclure la collecte, notamment lorsque les sources des données sont invérifiables voire illégales.

EN RÉSUMÉ ET PAR DÉFINITION…

Les critères dits « intrinsèques et essentiels » permettent d'évaluer la qualité des données en elles-mêmes, indépendamment du contexte d'usage ou de leur gouvernance.

Les critères « contextuels et contributifs » ont trait à la pertinence des données dans leur environnement d'usage, ainsi qu'à leur gouvernance, à la sécurité et à la capacité opérationnelle de l'entreprise.

NEUF CRITÈRES INTRINSÈQUES ET ESSENTIELS DE MESURE

CRITÈRE ET SYNONYMES	DESCRIPTION COURTE	EXEMPLES CONCRETS	INDICATEURS DE MESURE (KPI)
Complétude/ Exhaustivité	Tous les champs et enregistrements nécessaires sont présents pour l'usage prévu.	▸ Dossier client avec nom, adresse, email, téléphone et consentement RGPD renseignés. ▸ Fiche produit e-commerce avec images, dimensions, poids et catégorie.	Pourcentage de champs obligatoires remplis ; pourcentage d'enregistrements complets par entité.
Unicité/ Non-duplication	Chaque entité réelle est représentée une seule fois dans le jeu de données.	▸ Un client ne figure qu'une fois malgré les variations d'orthographe.	Taux de doublons.
Exactitude/ Justesse	Conformité à la réalité : la valeur observée est correcte.	▸ Le code SIRET correspond à l'entreprise décrite. ▸ Le tarif facturé correspond au contrat en vigueur.	Pourcentage de valeurs vérifiées correctes ; écart moyen vs source de référence ; taux d'erreurs détectées par audit.
Validité/ Conformité aux règles	Respect des formats, types, domaines et contraintes métier.	▸ Email conforme (syntaxe et domaine valide). ▸ Date de commande ≥ date d'ouverture du compte.	Pourcentage de valeurs passant les règles de validation ; violations de domaine/format ; taux de rejets aux contrôles.
Cohérence/ Intégrité/ Vraisemblance/ Logique interne	Absence de contradictions, respect des relations (intégrité référentielle) et plausibilité des valeurs.	▸ Client « actif » sans date de résiliation passée. ▸ Commande relative à un client existant (clé étrangère valide)[1].	Nombre de violations de contraintes ; pourcentage de clés étrangères valides ; taux de valeurs improbables (ex. âge > 120).
Fiabilité/ Crédibilité	Confiance dans la donnée et sa stabilité (source réputée, processus contrôlé).	▸ Données issues d'un référentiel. ▸ Mesures d'un capteur calibré avec contrôles périodiques.	Pourcentage de données provenant de sources à haut niveau de confiance ; taux d'anomalies récurrentes ; score de réputation de la source.
Actualité/ Fraîcheur	La récence des données par rapport au cycle attendu (SLA de mise à jour[2]).	▸ Stock mis à jour dans l'heure. ▸ Statut de livraison rafraîchi en temps réel.	Âge moyen des données ; pourcentage de mises à jour dans la fenêtre SLA ; latence de rafraîchissement.
Précision/ Granularité	Niveau de détail ou de résolution de la donnée.	▸ Coordonnées GPS avec 6 décimales (résolution fine). ▸ Montants financiers au centime plutôt qu'à l'euro entier.	Nombre de décimales ; distribution des intervalles ; couverture des niveaux de détail requis.
Uniformité/ Homogénéité	Uniformité des formats, unités et codifications dans le jeu de données.	▸ Toutes les dates en ISO 8601 UTC. ▸ Prix exprimés uniquement en EUR (pas de mélange de devises).	Pourcentage de valeurs au format standard ; pourcentage d'unités harmonisées ; nombre de variantes détectées par champ.

[1] Une clé étrangère est une valeur dans une table qui doit correspondre à une valeur existante dans une autre table. Elle permet de relier les données entre elles et de garantir la cohérence des informations.

[2] Un SLA de mise à jour définit le temps maximal acceptable entre la production d'une information et sa mise à jour dans le système. Il permet de garantir que les données restent récentes, fiables et utilisables. Si une donnée est trop ancienne, elle n'a plus de valeur.

HUIT CRITÈRES CONTEXTUELS ET CONTRIBUTIFS

CRITÈRE ET SYNONYMES	DESCRIPTION COURTE	EXEMPLES CONCRETS	INDICATEURS DE MESURE (KPI)
Traçabilité/ Auditabilité	Capacité à retracer l'origine, les transformations et les accès.	▸ Journal d'audit (qui/quand/quoi) complet et immuable.	Pourcentage d'opérations journalisées.
Clarté/ Compréhensibilité des données/ Lisibilité	Facilité de compréhension (noms, définitions, métadonnées, conventions).	▸ Dictionnaire de données à jour, accessible. ▸ Nommage explicite des champs.	Pourcentage de champs documentés.
Utilité/ Pertinence/ Adéquation métier	Aptitude à l'usage et valeur pour les processus et décisions.	▸ Indicateurs réellement utilisés dans la prise de décision. ▸ Attributs indispensables au processus (ex. taille/poids pour la logistique).	Taux d'utilisation des champs ; contribution aux indicateurs métier ; pourcentage de données nécessaires à un processus critique.
Accessibilité	Facilité et rapidité d'accès pour les utilisateurs autorisés (interfaces, rôles).	▸ API documentée et stable.	Temps moyen d'accès ; taux de requêtes réussies ; score d'expérience utilisateur (UX).
Conformité/ Légalité/ Respect des normes	Respect des règlementations et des normes applicables.	▸ RGPD : minimisation, consentement, droits des personnes. ▸ Normes ISO (ex. 27001) et politiques internes applicables.	Nombre de non-conformités ; pourcentage d'exigences satisfaites ; résultats et remédiations d'audits.
Sécurité/ Protection	Confidentialité et intégrité des données, contrôle d'accès, chiffrement.	▸ Chiffrement au repos et en transit.	Nombre d'incidents ; couverture de chiffrement.
Disponibilité	Capacité à accéder aux données quand nécessaire.	▸ Requêtes servies sans time-out pendant des pics de charge.	Taux de requêtes servies dans les délais.
Pérennité/ Durabilité	Accessibilité et lisibilité sur le long terme.	▸ Archivage en format pérenne (CSV, PDF/A, Parquet). ▸ Migration de données avec tests d'intégrité et compatibilité.	Pourcentage de données en formats standards pérennes ; taux de succès des migrations ; couverture des politiques de conservation.

DE LA QUALITÉ À LA GOUVERNANCE DES DONNÉES

Nous venons de voir, **les neuf critères essentiels** de la qualité des données constituent **le socle opérationnel indispensable** pour évaluer et améliorer la performance d'un système d'information. En nous procurant une mesure précise et objective de l'état réel des données, ils nous permettent d'appliquer les corrections nécessaires.

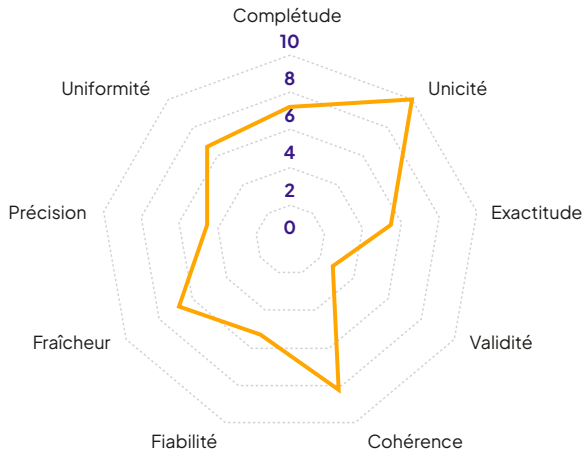
Mais ces critères, aussi fondamentaux soient-ils, ne suffisent pas à rendre compte de la qualité des données dans leur globalité. Ils doivent être **complétés par une vision contextuelle**. D'où l'importance des huit critères dits « **contextuels et contributifs** », qui vont au-delà de la logique purement technique pour intégrer la manière dont les données sont produites, transformées, utilisées et pilotées.

Et c'est précisément là que se situe **le point de basculement de la qualité** vers la gouvernance des données (ou « data gouvernance » en anglais).

La raison est que ces critères contributifs ne relèvent pas uniquement d'outils ou de contrôles automatisés mais bien de **règles**, de **rôles**, de **processus** et d'**arbitrages** qui encadrent la donnée. Leur qualité est par conséquent principalement le **résultat direct d'un cadre de gouvernance maîtrisé, partagé** et **appliqué** à l'échelle de toute l'entreprise.

Vous l'aurez compris : la qualité de données ne peut être pleinement efficace sans **une gouvernance à même d'en garantir la cohérence**, la pérennité et la responsabilité, un sujet central que la troisième partie de ce livre blanc vous permettra d'approfondir. En posant les **règles d'organisation, de structuration et de sécurisation** de l'ensemble du cycle de vie des données, la gouvernance est le véritable moteur d'une qualité durable.

Le graphique ci-dessous illustre la manière dont les neuf critères essentiels de qualité peuvent être visualisés sous forme de radar. Il ne s'agit pas ici de résultats réels ni de valeurs opérationnelles, mais d'un exemple destiné à montrer comment les indicateurs de mesure de la qualité peuvent être restitués **de façon synthétique et immédiatement lisible.**



Cette représentation offre une vue panoramique du niveau de qualité d'un ensemble structuré de données regroupées dans un format spécifique (dataset), d'une base de données complète ou même d'un système d'information. Elle permet d'identifier en un coup d'œil les **dimensions fortes**, les **faiblesses potentielles** et les **axes prioritaires d'amélioration**, facilitant ainsi le pilotage continu de la data quality.

2. Comment mesurer la qualité de vos données ?

Pour obtenir une vision claire de l'état de vos données, vous pouvez vous baser sur les KPI qui permettent de contrôler chaque critère de qualité. Toute la difficulté réside dans la diversité des données et dans l'identification du bon niveau d'indicateurs à suivre. Nous vous en donnons quelques exemples dans le tableau ci-dessous.
Une mesure objective est le point de départ indispensable et structurant pour la suite, qui consistera pour vous à décider des actions correctrices adéquates à mettre en œuvre.

TYPE DE DONNÉE **Adresse postale**	QUALITÉ SYNTAXIQUE	DÉLIVRABILITÉ
	▸ Mesure de la qualité de l'adresse RNVP (Restructuration, Normalisation et Validation Postale). ▸ Taux d'adresses validées aux normes postales.	▸ Utilisation de l'adresse dans le cadre d'une campagne postale. ▸ Taux de plis non distribués (PND) par La Poste.
RÉCENCE	**TAUX DE RÉACTIVITÉ**	**FIABILITÉ/ASSOCIATION**
▸ Récence de collecte, de mise à jour ou d'utilisation du contact. ▸ Date de la collecte ou de la mise à jour. ▸ Date du dernier envoi à l'adresse postale sans PND.	▸ Qualité du contact en termes de réactivité. ▸ Analyse des retours. ▸ Taux de transformation.	▸ La personne habite toujours à cette adresse. ▸ Risque de mauvaise attribution à une zone de chalandise ou à un profil du contact. ▸ Taux Estocade. ▸ Taux de doublons.

TYPE DE DONNÉE **Téléphone**	QUALITÉ SYNTAXIQUE	DÉLIVRABILITÉ
	▸ Numérique, format international (+33) ou français (06), nombre de caractères. ▸ Filtre mobile. ▸ Numéros DOM-TOM et étranger. ▸ Numéros fantaisistes (000000000). ▸ Préfixes non affectés. ▸ Taux de téléphones validés.	▸ Utilisation du téléphone en télémarketing ou pour le mobile en SMS/RCS/VMS. ▸ Taux hors faux numéros (avec tonalité). ▸ Taux d'aboutis pour SMS/RCS/VMS. ▸ Taux capability check (RCS).
RÉCENCE	**TAUX DE RÉACTIVITÉ**	**FIABILITÉ/ASSOCIATION**
▸ Récence de collecte, de mise à jour ou d'utilisation du contact. ▸ Date de la collecte ou de la mise à jour. ▸ Date du dernier routage abouti ou clic pour le SMS ou le RCS.	▸ Qualité du contact en termes de réactivité. ▸ Taux de transformation. ▸ Taux de décrochés. ▸ Taux de clics. ▸ Taux de vues (RCS)	▸ Du fait de la réaffectation des numéros de mobile, 10 % à 20 % d'une base de numéros de téléphone peuvent être affectés aux mauvaises personnes. ▸ Le numéro est abouti, mais le patronyme et l'adresse postale ne correspondent pas. ▸ Taux d'adresses validées via API KYC.

TYPE DE DONNÉE **Email**	QUALITÉ SYNTAXIQUE	DÉLIVRABILITÉ
	▸ Identifiant. ▸ Nom de domaine. ▸ Caractères spéciaux. ▸ Extension de domaine. ▸ Domaines poubelles. ▸ Taux d'emails validés.	▸ Routage des campagnes. ▸ Taux d'aboutis. ▸ Taux de hard bounces.
RÉCENCE	**TAUX DE RÉACTIVITÉ**	**FIABILITÉ/ASSOCIATION**
▸ Récence de collecte, de mise à jour ou d'utilisation du contact. ▸ Date de collecte ou de mise à jour. ▸ Date du dernier routage abouti, d'ouverture ou clic.	▸ Qualité du contact en termes de réactivité. ▸ Taux de transformation. ▸ Taux de clic. ▸ Taux d'ouvreurs.	

TYPE DE DONNÉE **Date de naissance**	QUALITÉ SYNTAXIQUE	DÉLIVRABILITÉ
	▸ Normalisation européenne JJ/MM/AAAA. ▸ Dates aberrantes ou issues de formulaires pré-remplis. ▸ 01/01/1900. ▸ Âge > 120. ▸ Taux de dates de naissance (DDN) validées.	
RÉCENCE	**TAUX DE RÉACTIVITÉ**	**FIABILITÉ/ASSOCIATION**
▸ Récence de collecte et de mise à jour. ▸ Date de collecte et de mise à jour.		▸ Analyse de la cohérence : fréquence des dates de naissance dans un fichier.

TYPE DE DONNÉE **Patronyme Nom/Prénom**	QUALITÉ SYNTAXIQUE	DÉLIVRABILITÉ
	▸ Caractères spéciaux. ▸ Nombre de caractères. ▸ Patronymes fantaisistes (ex. Spider Man). ▸ Existe dans les dictionnaires de Nom et Prénom. ▸ Taux de patronymes conformes.	
RÉCENCE	**TAUX DE RÉACTIVITÉ**	**FIABILITÉ/ASSOCIATION**
▸ Récence de collecte et de mise à jour. ▸ Date de la collecte et de la mise à jour.		▸ Comparaison avec une base tierce. ▸ Taux d'individus validés *via* API KYC.

TYPE DE DONNÉE BtoB	QUALITÉ SYNTAXIQUE	DÉLIVRABILITÉ
	▸ Société présente dans le référentiel Siren. ▸ Taux de « Sirétisation ».	
RÉCENCE	**TAUX DE RÉACTIVITÉ**	**FIABILITÉ/ASSOCIATION**
▸ Récence de collecte et de mise à jour. ▸ Société active. ▸ Taux d'activité (vs entreprise fermée).		▸ Contact travaillant toujours dans l'entreprise. ▸ Taux d'individus validés *via* API KYB.

TYPE DE DONNÉE ID device (cookie, ICTV, ID publicitaire mobile…)	QUALITÉ SYNTAXIQUE	DÉLIVRABILITÉ
		▸ ID activable dans une DSP ou autre plateforme.
RÉCENCE	**TAUX DE RÉACTIVITÉ**	**FIABILITÉ/ASSOCIATION**
▸ Récence d'activation ou de collecte.	▸ Taux de clics.	▸ Association avec un individu (onboarding).

3. Le cycle de vie de la donnée et ses points de contrôle

Gérer la qualité des données, c'est intervenir à plusieurs moments clés de leur cycle de vie. À chaque étape — création, stockage, utilisation ou simple conservation —, des contrôles spécifiques permettent de prévenir les erreurs, d'éviter la propagation d'anomalies et de maintenir un niveau de qualité durable.

- **Dès la création :** mettez en place des contrôles en temps réel pour une « data quality by design ».
- **Lors du stockage :** auditez la qualité de vos données lors de leur importation ou sur le stock existant.
- **Au moment de l'utilisation :** validez un jeu de données juste avant une campagne ou son utilisation dans un projet.
- **Dans la durée :** assurez-vous que les données restent à jour et conformes légalement au fil du temps, même si elles ne sont pas activement utilisées.

La gestion de la qualité des données et la mise en place de contrôles en temps réel nécessitent une approche structurée, des outils adaptés et des pratiques éprouvées pour garantir leur intégrité et leur fiabilité tout au long de leur cycle de vie.

DATA QUALITY BY DESIGN

Une démarche de « data quality by design » consiste à intégrer les processus de qualité dès la création des données, plutôt qu'à les corriger après coup.
Voici nos cinq recommandations pour vous assurer de correctement déployer une qualité de données « by design » :

a. Mettez en place une surveillance continue des données et leur validation dès leur création

La mise en place d'une surveillance continue des données permet de détecter instantanément les anomalies, le cas échéant, et de procéder à leur validation dès l'ingestion de nouveaux flux. Les systèmes d'observabilité et d'alerte intelligente bloquent l'enregistrement de données corrompues et redirigent immédiatement les incidents vers les responsables appropriés, à savoir le *chief data officer* ou le responsable des données (*data owner*[3]).

b. Appliquez des règles métier centralisées et automatisées

Définissez et appliquez des règles métier cohérentes pour tous les contrôles de qualité, puis automatisez leur exécution à travers les différents points d'entrée des données (ERP, CRM, fichiers…). C'est ce qui vous permettra de garantir la conformité et l'intégrité de vos données, quel qu'en soit la source. Les outils modernes proposent la réutilisation des règles dans divers environnements et la supervision centralisée des résultats.

c. Procédez à la vérification en temps réel avant le stockage

La validation de la qualité des données avant qu'elles ne soient enregistrées ou intégrées dans les bases opérationnelles garantit qu'aucune information erronée n'est propagée dans le système. Cela implique des contrôles instantanés (profilage, validation de formats, normalisation) dès la création ou la modification d'un enregistrement.

d. Déployez des outils de gouvernance et de gestion des données

Utilisez des plateformes de Master Data Management (MDM). Ces dernières aident à :

- **compiler** une version unique, qualitative et traçable de vos données ;
- **supprimer** les doublons ;
- **assurer** de manière automatisée le suivi et la correction des erreurs tout au long de leur cycle de vie ;
- **assurer** que les données de référence ayant démontré leur valeur stratégique soient diffusées dans les différents outils, désenclavant ainsi les silos.

e. Bénéficiez de l'automatisation et des retours en temps réel grâce à l'IA

L'automatisation des contrôles et la génération de rapports en temps réel, grâce à l'intégration de l'intelligence artificielle dans les outils de gestion de la qualité, offrent performance et adaptabilité. Ces systèmes fournissent également des indicateurs sur la santé des données et permettent des ajustements dynamiques sans intervention manuelle.

[3] Le *data owner* (en français « responsable des données ») est la personne responsable, côté métier, de la définition des règles d'utilisation, de qualité et d'accès pour un ensemble de données.

LA QUALITÉ DES DONNÉES LORS DE LEUR STOCKAGE

Au moment de l'importation des données ou durant leur stockage, l'audit de la qualité des données doit suivre des étapes structurées combinant analyse technique et métier.

Voici nos cinq recommandations pour vous permettre d'auditer efficacement vos données.

a. Mettez en place un profilage et un diagnostic initial des données importées

Avant même de les intégrer dans le système souhaité, il est crucial de réaliser un profilage des données sources afin d'identifier les anomalies, incohérences ou valeurs manquantes. Exemple de question à se poser : « *Mon fichier provient-il d'une source fiable, moyennement fiable ou peu fiable ?* »
Ce bilan offre une vue d'ensemble de la qualité de départ et oriente les directives commerciales ou automatisées de vérification à mettre en place.

b. Définissez les règles de validation durant l'importation

En mettant en place des règles métier précises, appliquées automatiquement lors de l'importation, vous vous assurez que seules les données conformes entrent dans le système.
Les outils de qualité offrent des fonctionnalités de validation, de nettoyage et de normalisation en temps réel.

c. Déployez des comparaisons et des contrôles croisés après les importations

Une fois que les données sont injectées et intégrées, il vous faut comparer les données sources et celles stockées afin de vous assurer qu'aucune information n'a été altérée ou perdue au cours de l'intégration. Ce contrôle croisé porte sur la complétude, la conformité et l'intégrité des champs sensibles.

d. Réalisez des audits et un suivi régulier des processus de traitements et de stockage

Il est crucial de procéder à des évaluations périodiques des procédures de traitement (avant, pendant et après le stockage). Ces audits permettent de s'assurer de la pérennité des processus, de déceler d'éventuels déclins et de maintenir une conformité permanente vis-à-vis des impératifs commerciaux et législatifs. Il est recommandé d'établir un plan de surveillance périodique (trimestriel ou semestriel) pour s'assurer du bon déroulé des processus.

e. Servez-vous de tableaux de bord et de journaux d'alertes

La mise en place de tableaux de bord de qualité (avec par exemple, des indicateurs sur les doublons, sur les valeurs manquantes ou le taux d'erreur) et de journaux d'alertes sur des erreurs automatiquement renseignées lors de chaque importation ou tentative de modification facilitera :

- le **suivi**,
- la **correction**,
- et la **traçabilité** des incidents.

Cela offre également un retour rapide vers les équipes métiers, afin de procéder aux corrections.

Ces pratiques vous procurent la maîtrise du cycle de vie des données, réduisent le risque d'incidents et garantissent de meilleures bases pour leur exploitation future.

LA QUALITÉ DES DONNÉES AU MOMENT DE LEUR UTILISATION

Même lorsque la validation des données est faite au fil de l'eau, lors de leur utilisation, il est important d'appliquer des règles claires. Ces dernières doivent être respectées avant toute activation marketing. Il est notamment hautement recommandé de :
évaluer le/les fichier(s) de données (format, complétude…) ;
- **nettoyer** les contacts (format, doublons…) ;
- **vérifier** la pertinence des données (dates de validation) ;
- **vérifier** les opt-in/opt-out (conformité) ;
- **réaliser** un test sur un échantillon.

Concernant l'usage dans le cadre d'un projet d'IA pour alimenter, contextualiser et entraîner un modèle, l'objectif pour vous ne sera plus de surveiller la donnée « en général », mais de vérifier qu'elle est suffisamment fiable pour un usage opérationnel immédiat.

Assurez-vous que les données sont cohérentes en :
- **évaluant** le fichier de données ;
- **réalisant** un profilage avec des champs (labels) ;
- **comparant** les données avec la segmentation définie ;
- **réalisant** une validation technique par des jeux de données.

Nous vous recommandons de réaliser deux checklists pour chaque usage avec dix éléments clés.

LA QUALITÉ DES DONNÉES DANS LA DURÉE

La gestion de la qualité des données dans la durée est un processus continu, structuré par une gouvernance solide et rythmée par des contrôles planifiés.

Nous reprenons ici en résumé les réflexes que vous devez avoir pour vous assurer de la qualité de vos données sur la durée :
- **Mettez en place la gouvernance de vos données** en définissant des règles de qualité globale.
- **Incarnez la qualité** en nommant des « ambassadeurs » de la qualité des données, des représentants data (DPO)… Vérifiez périodiquement vos données en mettant en place des contrôles réguliers.
- **Validez vos données** en suivant les KPI de contrôle.
- **Pensez à mettre en place des règles** qui empêchent vos données de « vieillir ». Les problèmes de qualité couramment observés dans les organisations sont aussi bien le fruit de données nouvelles mal contrôlées que du vieillissement progressif des données existantes.
- **Servez-vous de l'automatisation** en mettant à disposition de vos équipes des outils fiables et pertinents.

4. Le cadre règlementaire de la qualité des données et ses impacts

La qualité des données n'est plus seulement un sujet technique ou marketing : elle est devenue un enjeu **juridique**, **éthique** et une **question de réputation**. Le cadre règlementaire est un élément indissociable de la qualité des données.

Les règlementations européennes ont fait évoluer la notion de qualité de données vers une exigence de conformité, fondée sur :
- la **licéité** de la collecte,
- la **fiabilité** des données traitées,
- la **traçabilité** de leur cycle de vie,
- et la **responsabilité** des acteurs impliqués.

Une donnée de mauvaise qualité est désormais aussi une donnée non conforme.

LE RGPD : SOCLE CENTRAL DU CADRE RÈGLEMENTAIRE

Nous passons en revue ici les principes clés du **Règlement Général sur la Protection des Données** (RGPD) traitant tout particulièrement de la qualité des données.

Le **RGPD** établit plusieurs principes structurants qui définissent implicitement ce qu'est une donnée de qualité. Les voici :

Exactitude :
- Les données personnelles doivent être exactes et, si nécessaire, tenues à jour.
- Une donnée erronée ou obsolète est une non-conformité.

Limitation des finalités :
- Les données doivent être collectées pour des finalités déterminées, explicites et légitimes.
- La qualité s'évalue aussi par l'adéquation donnée/usage.

Minimisation des données :
- Les données doivent être pertinentes et limitées à ce qui est nécessaire.
- Trop de données = baisse de qualité et augmentation du risque.

Limitation de la conservation :
- Les données ne doivent pas être conservées au-delà de ce qui est nécessaire.
- Une donnée trop ancienne est une donnée dégradée.

Intégrité et confidentialité :
- La qualité inclut la protection contre l'altération, la perte ou l'accès non autorisé.

DROITS DES PERSONNES ET EXIGENCES DE QUALITÉ

Le RGPD renforce la qualité des données à travers **les droits des personnes** :
- droit de **rectification**,
- droit à **l'effacement**,
- droit à la **limitation du traitement**,
- droit à la **portabilité**.

Cela impose que :
- les chaînes de données soient **maitrisées**,
- que les référentiels soient **cohérents**,
- l'on soit capable de **corriger** ou **supprimer** une donnée partout où elle est répliquée.

RESPONSABILITÉS

Le RGPD distingue :
- les **responsable(s)** du traitement,
- les **sous-traitant(s)**.

Une donnée de mauvaise qualité générée par un partenaire engage l'ensemble de l'écosystème.

TRAÇABILITÉ ET GOUVERNANCE

Les exigences règlementaires impliquent de disposer de :
- la **cartographie** des flux de données,
- la **documentation** des traitements (registre RGPD),
- la **traçabilité** des transformations de données.

La qualité devient un objet de gouvernance, pas seulement un KPI technique.

AUTORITÉS DE CONTRÔLE ET LIGNES DIRECTRICES

Les autorités comme la CNIL publient régulièrement :
- des **recommandations**,
- des **guides de bonnes pratiques**,
- des **sanctions exemplaires**.

Ces décisions font de la qualité des données un levier de conformité, mais aussi de différenciation.

IMPACTS CONCRETS DU CADRE RÈGLEMENTAIRE

Impacts opérationnels
- Nécessité de disposer de **standards communs** de qualité de données.
- Harmonisation des **définitions**, **formats** et **règles** de gestion.
- Mise en place de **processus de contrôle** et **d'audit**.
- Gestion coordonnée des **incidents de données**.

Impacts stratégiques

- **Meilleure valorisation** des données dans un cadre sécurisé.
- **Réduction des risques** juridiques.
- **Avantage concurrentiel** par la qualité et la conformité.

La qualité des données n'est plus une option technique. C'est une exigence règlementaire, un facteur de confiance et un actif stratégique.

5. Pseudonymisation et anonymisation des données

L'anonymisation est un processus qui aboutit à la suppression irréversible des informations personnelles d'un enregistrement afin de ne plus permettre d'identifier l'individu correspondant aux autres données.

La pseudonymisation consiste quant à elle à remplacer des informations d'identification d'un enregistrement par un identifiant ou pseudo ne permettant qu'une identification indirecte de l'individu à l'aide de la correspondance pseudo/identité, elle-même stockée à part.

Prenons le cas d'enregistrements contenant nom, prénom, adresse, téléphone, email, date de naissance et antécédents médicaux.

La pseudonymisation consiste à séparer les informations personnelles des antécédents médicaux en gardant un identifiant commun. Une partie des informations personnelles (âge, département) peut accompagner les antécédents médicaux dans la mesure où ils sont insuffisants pour remonter jusqu'aux personnes.

Il faut noter que la présence de plusieurs critères (âge, code postal, profession...) peut être suffisante pour identifier la personne. Les termes ainsi posés « un avocat de sexe féminin de 53 ans habitant le 92340 » suffisent pour identifier quasi surement la personne dont il s'agit. On ne sera donc pas là face à une anonymisation. De même, une adresse ôtée de sa partie non nominative identifie précisément un foyer dans le cas d'une adresse « horizontale » (non collective).

CE QUE DIT LA RÉGLEMENTATION

Le RGPD **protège les individus identifiables** par leurs données personnelles (toute donnée non anonymisée).

Cela implique que l'anonymisation doit impérativement faire disparaître irrémédiablement la partie personnelle, afin que ce ne soit pas possible de la reconstituer avec des moyens raisonnables.

À noter qu'une entreprise B recevant des données pseudonymisées d'une entreprise A disposera en effet de données anonymisées tant qu'elle ne disposera d'aucun lien lui permettant de remonter aux données personnelles (jurisprudence de la Cour de justice de l'Union européenne).

ASPECTS OPÉRATIONNELS

Parmi les techniques permettant d'opérer ces traitements, nous pouvons citer la signature ou hachage, qui consiste à passer les données dans une fonction qui les transforme de manière déterministe et en principe non inversible (ex: MD5, SHA2-256). Elle sert, par exemple, à vérifier un mot de passe. Le système garde le résultat de la signature du mot de passe, ce qui permet de vérifier lorsqu'il est saisi en clair de recalculer et vérifier la correspondance des signatures sans jamais le stocker. La seule attaque avérée à ce type de procédé est l'attaque dite « par dictionnaire », c'est-à-dire le calcul à l'avance de tous les cas possibles afin d'indexer tous les résultats. Mais ce type d'attaque est irréaliste lorsque le nombre de calculs dépasse les capacités conventionnelles des machines[4] .

Correctement utilisé, ce type d'algorithme respecte les critères du RGPD pour considérer une information comme anonyme en signant les données personnelles, tout en permettant de comparer deux enregistrements pour vérifier qu'il s'agit bien de la même personne sans toutefois l'identifier.

IMPACTS DE L'INFORMATIQUE QUANTIQUE

Les technologies de l'informatique quantique et leurs algorithmes[5] remettent en cause la sécurité de certaines pratiques. Même en l'absence d'ordinateurs quantiques opérationnels à ce jour, il est tout à fait légitime de questionner la pérennité des techniques d'anonymisation actuelles. Qu'adviendra-t-il des algorithmes utilisés actuellement lorsque l'ordinateur quantique sera devenu une réalité (si tant est qu'il existe un jour !) ?

Sur le plan théorique, les algorithmes de chiffrement (DSA, RSA, https, la sécurité des certificats...) devront être « durcis », par exemple par un doublement du nombre de bits pour la taille de clés. Ceci dit, dans le cas spécifique des algorithmes de signature utilisés pour la pseudonymisation, ces derniers ne seront pas fondamentalement remis en cause par le quantique. Ce qui menace leur pérennité, c'est surtout les progrès des techniques de déchiffrement et l'amélioration des capacités des ordinateurs conventionnels, comme ceux qui déprécient périodiquement les algorithmes (cas du MD5).

[4] Des organismes comme la NSA qui détiennent d'énormes capacités informatiques sortent du champ du raisonnable au sens du RGPD. Algorithmes de Shor et de Grover.

[5] Algorithmes de Shor et de Grover.

6. Standards de qualité et d'usage des données non personnelles

QU'EST-CE QU'UNE DONNÉE NON PERSONNELLE ?

Une donnée électronique non personnelle est **une donnée qui ne se rapporte pas à une personne physique identifiée ou identifiable** (article 2 du règlement européen 2018/1807 du 14 novembre 2018). Il y a deux cas de figure : soit la donnée non personnelle **n'a jamais concerné une personne physique** ; soit elle a été **anonymisée**.
En pratique, beaucoup de jeux de données « non personnelles » sont en réalité **mixtes** (personnelles + non personnelles). Parfois elles sont même difficilement séparables, ce qui impose d'appliquer le RGPD à l'ensemble.

En B2B, les données non personnelles sont :
- **Les données d'entreprise :** raison sociale, SIRET, secteur, chiffre d'affaires, taille, localisation, technologies utilisées, organigramme hiérarchique.
- **Les coordonnées génériques :** emails de type contact@, support@, info@, numéros standards, adresses postales du siège ou des agences.

En B2C, on peut citer :
- **Les données purement techniques :** volumes des ventes, volumes des ventes par produit, taux de clic global d'une campagne...
- **Les données transactionnelles non rattachées à un identifiant client individuel :** moyens de paiements, quantités vendues...
- **Les données statistiques sur les clients :** âge, géographique.
- **Les données d'enquêtes :** résultats agrégés...

CRITÈRES DE QUALITÉ DES DONNÉES NON PERSONNELLES

En fonction de l'usage qui est fait des données, différents aspects doivent être pris en compte. Tout d'abord, pour un usage à de fins d'analytique, la qualité sur :
- **l'exactitude** des données à jour,
- leur **complétude**,
- leur **traçabilité**,
- leur **gouvernance**.

Pour le marketing et la prospection, les standards d'usage visent à limiter les risques :
- **d'agrégation** et **d'anonymisation** de données sensibles, comme les données personnelles.
- **de biais**, pour éviter des pratiques de ciblage jugées discriminatoires.

Il est également recommandé de documenter les éventuels regroupements (segmentations) en indiquant les objectifs, les données utilisées et les périodes...

Enfin, pour un usage dans l'intelligence artificielle, et en particulier en machine learning, il n'y pas de secret : vu que le modèle apprend à partir des données qu'on lui fournit, si ces dernières sont incomplètes, biaisées, déséquilibrées ou incohérentes, le modèle reprendra, renforcera et amplifiera ces défauts. L'absence de qualité de données en IA est la porte ouverte vers des performances dégradées, des prédictions non conformes ou encore le manque de robustesse.

Voici quelques types de biais courant causés par des défauts de qualité dans les données utilisées en IA :

- **Biais d'échantillonnage :** les données ne représentent pas correctement la réalité.
- **Biais de labellisation :** pour un apprentissage supervisé, les étiquettes sont erronées, incohérentes ou subjectives.
- **Biais liés aux valeurs manquantes :** le modèle apprend à partir des données absentes pour certains profils ou encore les données absentes ne sont pas réparties au hasard.
- **Biais historiques :** l'IA reproduit des décisions passées à partir de données déjà biaisées.
- **Biais d'attribution :** les variables utilisées pour l'apprentissage sont aberrantes ou incohérentes.

Face à ces problèmes, **de nombreux outils statistiques** existent pour corriger les données : règles d'imputation, détection et traitement des valeurs aberrantes, gestion du déséquilibre des classes, normalisation et standardisation des variables, vérification de la cohérence et des dépendances.

La data gouvernance demeure elle aussi **indispensable pour assurer la qualité des données** dans l'IA. On peut citer l'importance de la traçabilité pour comprendre comment un modèle apprend, ou encore le rôle de la responsabilisation pour définir qui valide les données d'entrainement.

En définitive, la qualité des données n'est pas seulement un prérequis technique pour l'IA, elle en est la véritable garante qui conditionne la performance et la robustesse des modèles.

Partie 2.

Feuille de route pour une démarche de qualité des données réussie

1. L'importance de l'acculturation des équipes

En matière de qualité des données, parler uniquement d'outils, de processus ou de conformité serait une erreur.

Le véritable défi consiste à acclimater les équipes, et pas seulement celles de la DSI ou du marketing : toutes les directions d'une entreprise doivent être impliquées, vu que chaque service crée, transforme ou utilise des données.

La qualité de données est un enjeu transversal. Voici pourquoi.

LA DONNÉE N'APPARTIENT PAS À L'IT, ELLE APPARTIENT À TOUTE L'ENTREPRISE

Chaque direction produit, modifie ou utilise des données :
- Le **marketing collecte** des leads.
- Le **commercial enrichit** le CRM.
- Les **opérations traitent** les données de production, de logistique et de qualité.
- La **finance consolide** des données de facturation.
- Les **RH gèrent** des données sensibles des collaborateurs.
- Le **service client met à jour** des informations en continu.

Tant que chacun ne comprendra pas l'impact de ses actions sur la qualité globale, la donnée se dégradera mécaniquement et l'ensemble de l'entreprise sera impactée.
Sans acculturation, la data quality devient un « sujet technique ».
Avec l'acculturation, elle devient un enjeu stratégique partagé.

LA QUALITÉ DES DONNÉES EST AVANT TOUT LE FRUIT D'UN COMPORTEMENT HUMAIN

Les problèmes de qualité viennent rarement uniquement des outils.
Ils viennent surtout :
- de mauvaises pratiques,
- d'un manque de rigueur,
- de l'absence de standards,
- d'un défaut de compréhension des enjeux,
- des silos entre les directions.

La donnée est **interconnectée.** La qualité est par **conséquent collective**.

LA CONFORMITÉ RÈGLEMENTAIRE EXIGE UNE RESPONSABILITÉ COLLECTIVE

Depuis l'entrée en vigueur du RGPD, la gestion de la donnée ne peut plus se permettre d'être approximative.

Toutes les directions doivent comprendre qu'une donnée inexacte, obsolète ou mal tracée peut entraîner de graves conséquences pour l'organisation, à savoir :
- des **sanctions** administratives,
- des **amendes** importantes,
- des **contentieux** clients,
- des **atteintes à la réputation**,
- la **responsabilité civile** de l'entreprise.

Si les équipes ne sont pas acculturées :
- les **consentements** sont mal enregistrés,
- les **données** sont conservées trop longtemps,
- les **exports** sont réalisés sans contrôle,
- les **demandes de suppression** ne sont pas traitées correctement.

Le risque ne vient pas d'une mauvaise intention, mais d'un manque de compréhension.

Chaque collaborateur qui manipule des données engage potentiellement la responsabilité de l'entreprise.

L'acculturation permet de :
- **réduire le risque** d'erreur humaine,
- **ancrer des réflexes** de conformité,
- **sécuriser les processus** internes,
- **démontrer de la diligence** en cas de contrôle.

Dans un contexte de contrôles accrus et de durcissement des sanctions, la data quality devient un enjeu de gestion du risque juridique.

Sans acculturation, la conformité repose sur quelques personnes.
Avec l'acculturation, elle devient une culture d'entreprise.

L'ACCULTURATION CASSE LES SILOS ET AMÉLIORE LA GOUVERNANCE

La qualité des données nécessite que toutes les directions fonctionnent ensemble sur :
- des **référentiels** communs,
- des **définitions** partagées (qu'est-ce qu'un « client actif » ?),
- des **règles** de gestion harmonisées,
- des **responsabilités** claires (data owner, data steward[6]…).

Si chaque direction dispose de sa propre définition et de ses propres pratiques, la donnée devient incohérente.

L'acculturation permet d'aligner l'ensemble des directions autour d'un langage commun.

[6] Le data steward ("gestionnaire de données" en français) est la personne responsable de la gestion opérationnelle, de la qualité et de la documentation des données.

LA DONNÉE N'APPARTIENT PAS À L'IT, ELLE APPARTIENT À TOUTE L'ENTREPRISE

Aujourd'hui, la donnée – et par conséquent sa qualité – est un actif stratégique au sein d'une organisation, car :

- chaque direction **produit de la donnée**,
- toutes **prennent des décisions** basées sur la donnée,
- les risques **sont partagés**,
- la performance globale **dépend de la fiabilité** des informations.

Une entreprise qui comprend la valeur de sa data :

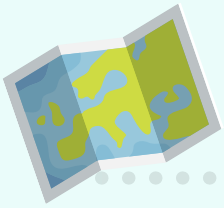
- **investit** mieux,
- **automatise** intelligemment,
- **personnalise** ses parcours,
- **prend des décisions** plus rapides.

Mais cette valeur n'existe que si la donnée est fiable, et la fiabilité repose sur la culture de la donnée.

La data quality est devenue un véritable enjeu stratégique pour tous les salariés d'une entreprise, quel que soit leur fonction.

2. Les étapes clés de votre projet

Une démarche complète de qualité de données peut se décomposer comme suit :



1. Audit & Cartographie

Cette première étape consiste à dresser un état des lieux et à cartographier les données et leurs flux (mapping). Vous devez y inclure toutes les étapes, de la collecte aux traitements réalisés et à l'exploitation des données. Veillez à bien préciser les noms des différents intervenants (de l'entreprise interne ou externes) impliqués ainsi que tous les outils déployés.

C'est aussi à cette étape que vous auditez la qualité de vos contacts : mesure de la qualité de l'adresse RNVP, délivrabilité email et SMS, taux de complétude, taux de doublons, etc.

2. Analyse des causes

Après avoir identifié ce qui ne fonctionne pas bien, il s'agit d'en analyser les causes en s'intéressant :

- à la modalité de collecte : formulaire web, magasin, télémarketing, partenaire, données obligatoires ou non lors de la collecte, etc ;
- aux traitements réalisés (ou non réalisés) ;
- aux règles de gestion de la suppression de doublons et/ou de la déduplication entre les bases ;
- à l'ancienneté de la collecte ou de la mise à jour des données.

3. Mise en place de mesures correctives

À chaque problème, sa solution avec des traitements de correction ponctuels ou périodiques. Nous vous en citons quelques-uns :

- suppression des doublons et déduplication des bases (référentiel client unique/RCU) ;
- RNVP – Restructuration, Normalisation et Validation Postale – des adresses ;
- identification des déménagés ou décédés ;
- capability check des mobiles ;
- mise en place de l'unicité des identifiants de navigation ;
- sirétisation des bases BtoB et validation de l'activité des entreprises ;
- vérification des identités clients (KYC) ;
- redressements syntaxiques des emails, dates de naissance, prénoms ou patronymes, etc.

Pensez toujours à épurer périodiquement vos bases de données : c'est indispensable à la fois pour des raisons de conformité et de qualité).

4. Prévention

Cette étape et clé et ambitieuse car elle consiste à faire évoluer son organisation et ses outils afin de prévenir tout risque de retour à une situation où la qualité ferait à nouveau défaut.

Cela commence par la définition des règles de gestion et des normes internes à même de garantir la qualité de la collecte et des traitements sur le long terme pour ensuite les transformer en nouvelles consignes à transmettre à toutes les équipes concernées par la data dans l'entreprise : commerciaux, call center, magasin, etc.

Voici quelques exemples de règles et outils de collecte à mettre en place :

- champs obligatoires (siret, date de naissance…) ;
- règles de gestion (nombre de caractères minimum du nom, syntaxe téléphone, email, code postal ;
- aide à la saisie ou API (auto-complétion des champs d'adresse ou des fiches entreprise) ;
- double authentification pour l'email et le mobile ;
- règles de référentiel client unique (RCU) pour l'intégration ou la déduplication entres sources ou outils différents ;
- API d'enrichissement des champs manquants (email, adresse postale, date de naissance) ;
- traçabilité des données (source, date de collecte et date de dernière activité).

5. Pilotage

Une fois la démarche de qualité de données établie et lancée, il s'agit de la piloter en suivant son évolution dans le temps à l'aide notamment de tableaux de bord.

Voici les indicateurs de qualité de base à suivre : taux de plis non distribués (PND), taux de déménagés, taux de délivrabilité email ou mobile, taux de complétude, récence moyenne, taux d'identification de visiteurs web, taux de « no cookie »…

3. Cas concrets et illustrations

- **Cas d'usage 1 :** Marketing - Fiabiliser une base de données clients avant le lancement d'une campagne.
- **Cas d'usage 2 :** Intelligence Artificielle - Préparer et nettoyer un dataset pour l'entraînement d'un modèle prédictif.
- **Cas d'usage 3 :** Opérationnel - Garantir la cohérence des données de référence (produits, fournisseurs) dans un système d'information.

CAS D'USAGE	EXEMPLES DE DYSFONCTIONNEMENTS	MESURES DE PRÉVENTION
Marketing - Fiabiliser une base de données clients avant le lancement d'une campagne. - Augmenter la performance des campagnes (ouverture, clic, conversion) en fiabilisant la base clients.	- Adresses email invalides ou obsolètes. - Doublons clients qui faussent les chiffres et irritent (plusieurs emails à la même personne). - Données incomplètes (segment, secteur, pays, langue) qui limitent la personnalisation. - Consentements RGPD mal tracés qui exposent à des risques légaux.	- Règles de saisie dans le CRM (champs obligatoires, listes de valeurs). - Intégration de connecteurs normalisés dans les CRM/outil de marketing pour éviter les imports manuels non contrôlés. - **Pilotage :** suivre dans un tableau de bord le taux d'emails invalides, de doublons, de fiches complètes, avant et après chaque campagne.

FEUILLE DE ROUTE DE LA QUALITÉ DE DONNÉES	RÉSULTATS ATTENDUS
- **Audit :** mesurer le taux d'emails invalides, de doublons, de fiches incomplètes. - **Analyse des causes :** formulaires avec champs non obligatoires, importations de fichiers non contrôlés. - **Correctif :** 1. Déduplication structurée (règles de rapprochement sur email, téléphone, nom et société). 2. Vérification syntaxique et validation des emails (opt-in de confirmation) si possible. 3. Campagnes de « data refresh » (formulaires pré-remplis pour que le client confirme ou corrige ses informations).	- Moins de rebonds, meilleure délivrabilité. - Meilleure segmentation, messages plus pertinents. - Reporting plus fiable (CA par segment, ROI campagne, etc.).

CAS D'USAGE	EXEMPLES DE DYSFONCTIONNEMENTS	MESURES DE PRÉVENTION
IA ▸ Préparation d'un dataset pour réaliser un modèle prédictif. ▸ Améliorer la performance et la robustesse d'un modèle (churn, scoring, recommandations) en garantissant la qualité du dataset.	▸ Valeurs manquantes sur des variables clés (revenu, ancienneté, usage produit). ▸ Incohérences de formats (dates, montants, devises, séparateurs). ▸ Variables avec beaucoup d'erreurs de saisie ou de libellés différents pour la même valeur. ▸ Données non représentatives (biais : surreprésentation d'un segment, historique partiel).	▸ **Renforcer la qualité à la source :** interfaces de saisie, contraintes sur les champs critiques. ▸ Mettre en place un processus entre producteurs et consommateurs de données → structure, format, champs sémantiques (ce que doit garantir chaque source pour être exploitable par les modèles). ▸ **Pilotage :** Suivre des indicateurs dédiés aux datasets de modèle.

FEUILLE DE ROUTE DE LA QUALITÉ DE DONNÉES	RÉSULTATS ATTENDUS
▸ **Audit :** analyser la complétude, la cohérence, la distribution des variables. ▸ **Analyse des causes :** • Systèmes sources qui ne collectent pas certaines informations. • Champs non obligatoires ou mal renseignés. • Processus métier qui n'imposent pas la mise à jour des données. ▸ **Correctif :** 1. Stratégies de traitement des valeurs manquantes (selon des règles métier). 2. Normalisation des formats (dates, devise). 3. Harmonisation des catégories (par exemple, fusionner FR, France, FRANCE en une valeur unique). 4. Détection et traitement des données non conformes ou dites aberrantes.	▸ Modèle plus performant, moins instable. ▸ Réduction des biais liés à des données incorrectes ou incomplètes. ▸ Capacité à ré-entraîner régulièrement le modèle sur des données contrôlées.

CAS D'USAGE	EXEMPLES DE DYSFONCTIONNEMENTS	MESURES DE PRÉVENTION
Opérationnel ▸ Assurer la cohérence et l'unicité des référentiels (produits, fournisseurs, clients) pour fiabiliser les opérations et le reporting.	▸ Un même produit présent avec plusieurs codes ou libellés dans différents systèmes. ▸ Informations fournisseurs différentes selon les applications (CRM, outils de BI, ERP, outil d'achats, logistique). ▸ Erreurs de classification (mauvaise famille de produit) qui perturbent les analyses. ▸ Création non conforme de nouvelles fiches produits ou fournisseurs sans vérification.	▸ Création d'un référentiel avec des règles de fusion et de suppression de doublons. ▸ **Correction des fiches existantes :** Alignement des codes, reclassement dans les bonnes familles. ▸ **Prévention :** Workflow de création/modification (demande, validation et publication) avec responsabilités claires. Règles de nommage, d'attribution de codes, de gestion des statuts (actif, obsolète, archivé). ▸ **Pilotage :** Indicateurs de qualité sur le référentiel avec le nombre de doublons, le taux de fiches complètes, le nombre de créations hors processus, le délai de création/validation d'un nouveau produit.

FEUILLE DE ROUTE DE LA QUALITÉ DE DONNÉES	RÉSULTATS ATTENDUS
▸ **Audit & cartographie :** • Identifier tous les référentiels existants, les champs clés (ID produit, ID fournisseur, famille, marque) et les flux entre systèmes. • Mesurer le nombre de doublons, d'incohérences, de fiches incomplètes. ▸ **Analyse des causes :** • Absence de référentiel maître ou de règles de création. • Processus métier décentralisés, chacun créant ses propres codes.	▸ Moins d'erreurs opérationnelles (commandes, livraisons, facturation). ▸ Reporting fiable sur les ventes, les marges, les volumes d'achat. ▸ Réduction des couts liés aux corrections manuelles et aux erreurs de référentiel

Partie 3.

De la qualité à la gouvernance des données

1. Le lien indissociable entre qualité et gouvernance des données

La qualité et la gouvernance des données sont étroitement liées et se renforcent mutuellement. **La qualité** des données vise à garantir que ces dernières sont **fiables**, **complètes**, **cohérentes** et **exploitables**. **La gouvernance**, elle fournit le cadre organisationnel pour atteindre et maintenir ce niveau de qualité dans le temps. Elle définit les **règles**, les **processus**, les **rôles** et les **responsabilités** nécessaires à l'assurance d'une utilisation maitrisée, cohérente et sécurisée des données au sein de l'entreprise.

Toute démarche de qualité des données révèle rapidement l'importance d'une gouvernance structurée. De fait, les outils ou les contrôles techniques ne changeraient pas grand-chose sans une organisation claire autour de la donnée, avec des responsabilités définies, des règles partagées et des processus formalisés.

Pour définir des règles de qualité, mettre en place des indicateurs et corriger des anomalies, il faut déjà pouvoir s'accorder sur des définitions communes, des usages partagés et des rôles clairement identifiés. C'est pourquoi toute initiative visant à renforcer la qualité des données conduit naturellement à instaurer des principes solides de gouvernance.

2. Pourquoi et pour quels bénéfices établir un cadre de gouvernance des données ?

Les organisations qui réussissent à industrialiser la donnée partagent une conviction : la qualité ne se décrète pas, elle se gouverne. Sans un cadre de gouvernance, les données se fragmentent, les définitions divergent, les processus se contredisent et la conformité se fragilise. À l'inverse, une gouvernance claire — c'est-à-dire un système de règles, de rôles, de responsabilités, et de décisions — permet de garantir la fiabilité des données tout en sécurisant leur usage et en maximisant leur valeur.

Les bénéfices que l'on retire d'une gouvernance des données sont nombreux. Les voici :

- **Instauration de la confiance :** des données traçables, claires, utiles et conformes sont acceptées par les métiers et les partenaires.
- **Accélération des projets :** moins de débats sémantiques, davantage de sources maitrisées, d'accès et des responsabilités définies réduisent les délais.
- **Diminution des risques :** une conformité règlementaire renforcée et une sécurité maîtrisée réduisent les risques de quelque nature qu'ils soient.
- **Amélioration du pilotage :** une cohérence transverse entre les directions finance, risque, marketing et opérations améliore les décisions à tous les niveaux.
- **Valorisation économique de la donnée :** les données sont réutilisables, capitalisées et peuvent même faire l'objet, dans certains cas, d'une monétisation.
- **Efficience opérationnelle :** moins de retraitements et moins de corrections, c'est aussi plus d'efficacité et moins de coûts.

3. Comment définir la gouvernance des données ?

Comme nous avons vu plus haut, la gouvernance des données est l'ensemble des principes, règles, rôles, processus et mécanismes de décision qui encadrent la manière dont une organisation définit, produit, partage, sécurise, qualifie et valorise ses données sur tout leur cycle de vie.

Elle vise un **triple alignement** :

- **Stratégique :** relier les priorités business aux politiques et aux investissements data.
- **Opérationnel :** orchestrer les responsabilités (qui décide quoi, qui fait quoi, avec quels outils).
- **Risque & Conformité :** se conformer aux cadres légaux et normatifs, protéger les actifs d'information.

La qualité des données est souvent un résultat observable et mesurable à travers les critères essentiels. La gouvernance est le cadre qui rend ce résultat prévisible, mesurable et durable.

La gouvernance des données est **transversale** par nature. Elle associe :

1. Les **métiers** (propriétaires et consommateurs de données), garants de l'utilité et de la clarté des concepts.
2. Le **data management** (data owners, data stewards, CDO), garant de la cohérence, de la documentation, des catalogues et des métriques. La gestion des données (data owners, data stewards, CDO) assure la conformité, la traçabilité, les répertoires et les indicateurs.
3. L'**IT/Architecture/Ops** (CIO, architectes, SRE), responsables de l'accessibilité, de la disponibilité, de la traçabilité technique et de la pérennité des environnements.
4. La **Sécurité/RSSI** et **Continuité (BCP/DRP)**, responsables de la sécurité, de la gestion des risques.
5. Le **Juridique/Conformité/DPO/Audit interne**, garants de la conformité et de la traçabilité probatoire.
6. La **direction générale** et le **risk management**, qui arbitrent, priorisent et sponsorisent.

Ce pilotage multifonctionnel se matérialise par des instances (Comité Data, Conseil de Gouvernance...) et des processus d'arbitrage (priorisation de cas d'usage, exceptions, niveaux de risque acceptables).

4. La transversalité des 8 critères contextuels et contributifs

Les huit critères contextuels et contributifs peuvent se répartir en quatre familles de gouvernance qui illustrent la transversalité da la gouvernance de données :

Famille 1

Gouvernance règlementaire & conformité

Conformité + Traçabilité

ACTEURS CLÉS

Juridique, DPO, Conformité, Audit interne, IT, Data

EXEMPLE :
Pour un reporting règlementaire, la traçabilité démontre l'origine des chiffres et la conformité garantit que seules les données nécessaires sont traitées et conservées selon les durées légales.

Famille 2

Gouvernance de la sécurité & de la résilience

Sécurité + Disponibilité

ACTEURS CLÉS

RSSI, IT, Risk management, Architecture

EXEMPLE :
Pour un service analytique critique, la disponibilité est contractualisée (SLA) et la sécurité impose un chiffrement et la redondance.

Famille 3

Gouvernance de l'usage & de la valeur

Utilité + Accessibilité + Clarté

ACTEURS CLÉS

Métiers, CDO, Product owners data, Data stewards, IT (interfaces & intégration)

EXEMPLE :
Pour le marketing, l'utilité s'évalue via le gain business vs coût d'alimentation ; l'accessibilité *via* le catalogue et des jeux de données « certifiés » ; la clarté via un glossaire partagé.

.

Famille 4

Gouvernance du cycle de vie & de la pérennité

Pérennité

ACTEURS CLÉS

Architecture, Records management, IT, Juridique

EXEMPLE :
Archivage probant à long terme : formats durables, métadonnées descriptives, horodatage, preuves d'intégrité, politiques de purge.

5. Une transformation organisationnelle

La mise en œuvre de la gouvernance des données peut **transformer l'entreprise à trois niveaux** :

1. **Culturel :** on passe des « données subies » aux actifs gérés. Les métiers deviennent coproducteurs de qualité, les équipes data deviennent facilitatrices de valeur, et la conformité n'est plus vécue comme un frein, mais comme un accélérateur de confiance.

2. **Organisationnel :** les rôles sont clarifiés, les arbitrages sont plus rapides, les duplications de sources diminuent, la production de rapports et de produits data devient prévisible et industrialisée.

3. **Économique :** la donnée est capitalisée — documentée, traçable, réutilisable — donc moins coûteuse à maintenir et plus rentable à exploiter. Les cas d'usage s'enchaînent plus vite, l'interopérabilité avec partenaires et régulateurs s'améliore, et la valeur créée (revenus additionnels, réduction de la fraude, optimisation des stocks, meilleure acquisition/retention) est mesurable.

En pratique, le cadre de la gouvernance allie **rigueur** (conformité, sécurité, pérennité) et **agilité** (utilité, accessibilité, clarté), orchestré par un **pilotage multifonctionnel**. C'est cette combinaison qui installe la confiance et permet de capitaliser sur la donnée : **moins d'efforts dispersés**, **davantage d'impact** et une **trajectoire durable** vers l'entreprise véritablement data-driven.

Conclusion

La qualité des données, un investissement continu sur l'avenir

Au fil de ce livre blanc, nous avons posé les **bases indispensables** pour comprendre, mesurer et piloter la qualité des données de manière durable.

Les critères d'appréciation de la qualité des données sont nombreux, mais leur répartition entre les **neuf critères intrinsèques** et les **huit critères contributifs** permet de mieux appréhender la complexité de ce sujet.

Nous avons souhaité vous donner des **exemples concrets** et de **nombreux angles d'analyse** à travers le cycle de vie de la donnée, le cadre règlementaire, la qualité appliquée aux données non personnelles et aux projets d'IA.

Nous avons même ouvert la porte vers le sujet de la **gouvernance des données**, dont l'importance pour la qualité des données n'est plus à démontrer.

Finalement, un enseignement majeur se dégage : **la qualité des données n'est pas un chantier ponctuel, mais un processus vivant, inscrit dans une logique d'amélioration continue. Elle doit évoluer avec les usages, les systèmes, les métiers et les règlementations.**

C'est en la traitant comme une discipline à part entière et non comme un projet technique qu'elle peut devenir un levier de **performance**, de **confiance** et de **gouvernance**.

Un dernier point : **la question de l'usage de l'IA pour améliorer la qualité des données ouvre des perspectives intéressantes.** Aujourd'hui, les professionnels s'accordent à reconnaître que les modèles actuels ne sont pas capables d'atteindre une fiabilité suffisante, les risques de dérive restant trop élevés.

Mais les avancées rapides de l'IA laissent penser que cette frontière pourrait évoluer et que l'IA pourrait devenir demain un **assistant puissant** pour détecter des incohérences, suggérer des corrections ou encore accélérer les traitements.

Bien sûr, l'IA devra être **encadrée**, **contrôlée** et **pilotée avec rigueur**. Une erreur de correction peut engendrer des conséquences plus néfastes que l'absence de correction, surtout lorsque l'on parle des données nominatives et/ou BtoB présentes dans les CRM.

Sans oublier la **gouvernance**, la **transparence** et la **maîtrise** qui restent, avec l'humain, le cœur du réacteur de ce dispositif.

MERCI AUX PARTICIPANTS

amabis
Donnons du sens à la data client

CBC DEVELOPPEMENT

ELLISPHERE
Trust. Powered by data

marketespace

Alliance Digitale
112ter rue Cardinet
75017 Paris
contact@alliancedigitale.org

alliance digitale : dma France MMAF iab france