

## CONCOURS INTERNE D'INGÉNIEUR TERRITORIAL

SESSION 2025

ÉPREUVE DE PROJET OU ÉTUDE

ÉPREUVE D'ADMISSIBILITÉ :

L'établissement d'un projet ou étude portant sur l'une des options, choisie par le candidat lors de son inscription, au sein de la spécialité dans laquelle il concourt.

Durée : 8 heures  
Coefficient : 7

**SPÉCIALITÉ : INFORMATIQUE ET SYSTÈMES D'INFORMATION**  
**OPTION : SYSTÈMES D'INFORMATION ET DE COMMUNICATION**

### À LIRE ATTENTIVEMENT AVANT DE TRAITER LE SUJET :

- ♦ Vous ne devez faire apparaître aucun signe distinctif dans votre copie, ni votre nom ou un nom fictif, ni initiales, ni votre numéro de convocation, ni le nom de votre collectivité employeur, de la commune où vous résidez ou du lieu de la salle d'examen où vous composez, ni nom de collectivité fictif non indiqué dans le sujet, ni signature ou paraphe.
- ♦ Sauf consignes particulières figurant dans le sujet, vous devez impérativement utiliser une seule et même couleur non effaçable pour écrire et/ou souligner. Seule l'encre noire ou l'encre bleue est autorisée. L'utilisation de plus d'une couleur, d'une couleur non autorisée, d'un surligneur pourra être considérée comme un signe distinctif.
- ♦ Le non-respect des règles ci-dessus peut entraîner l'annulation de la copie par le jury.
- ♦ Les feuilles de brouillon ne seront en aucun cas prises en compte.

**Ce sujet comprend 68 pages.**  
**Il appartient au candidat de vérifier que le document comprend le nombre de pages indiqué.**

*S'il est incomplet, en avertir le surveillant.*

- ♦ Vous répondrez aux questions suivantes dans l'ordre qui vous convient, en indiquant impérativement leur numéro.
- ♦ Vous répondrez aux questions à l'aide des documents et de vos connaissances.
- ♦ Des réponses rédigées sont attendues et peuvent être accompagnées si besoin de tableaux, graphiques, schémas...

Vous êtes ingénieur territorial, chef de projet à la direction des systèmes d'information de Commaglo, une communauté d'agglomération qui comprend 2 000 agents pour un territoire de 20 communes et 200 000 habitants.

Commaglo cherche à améliorer ses processus de gestion et de prise de décision grâce à une meilleure utilisation des données. Dans ce cadre, la collectivité a décidé d'intégrer des méthodes modernes de collecte, d'analyse et de visualisation des données pour répondre plus efficacement aux attentes de ses citoyens.

Le directeur des systèmes d'information (DSI) vous demande une étude en vue de mettre en place une solution (technique et organisationnelle) pour intégrer les méthodes de collecte, d'analyse et de visualisation des données. Il s'agit par ce biais d'optimiser la prise de décision au sein de Commaglo tout en garantissant la qualité, la sécurité et l'accessibilité des informations.

### **Question 1 (4 points)**

Vous rédigerez une note, à destination de la direction générale (DG), sur les enjeux et les impacts de l'évolution des méthodes de collecte, d'analyse et de visualisation des données. Vous préciserez également en quoi les données et leur interprétation peuvent jouer un rôle majeur dans la prise de décision pour Commaglo.

### **Question 2 (5 points)**

Vous détaillerez une démarche de projet pour l'intégration des méthodes de collecte, d'analyse et de visualisation des données. Les étapes seront représentées sous la forme d'un organigramme visuel.

### **Question 3 (3 points)**

Vous composerez une fiche technique, destinée à l'ensemble des encadrants de la Commaglo, pour :

- expliciter l'importance de la qualité des données pour une collectivité territoriale ;
- identifier les leviers permettant d'améliorer la qualité des données collectées.

#### Question 4 (4 points)

A la demande du DSI, et à son attention, vous rédigez une note technique détaillant en quoi les avancées en matière d'intelligence artificielle et de machine learning ont enrichi les méthodologies d'analyse des données. Vous préciserez également quels sont les défis associés à leur mise en œuvre.

#### Question 5 (4 points)

Vous formaliserez une proposition d'organisation et d'architecture du système d'information pour faire évoluer les méthodes de collecte, d'analyse et de visualisation des données dans les services de Commagglo.

#### Liste des documents :

**Document 1 :** « Le modern Data Lake va-t-il se généraliser ? » - *Journal du net* - 12 décembre 2023 - 2 pages

**Document 2 :** « Les 7 meilleurs outils de collecte de données en 2023 » - *questionpro.com* - consulté le 27 novembre 2024 - 8 pages

**Document 3 :** « Stratégie de gestion de données dans les collectivités territoriales : enjeux, implications et mise en œuvre » - *tbmaestro.com* - 23 avril 2024 - 4 pages

**Document 4 :** « Maîtrise des données territoriales : nouveaux enjeux, nouvelles solutions pour les collectivités » - *keyrus.com* - consulté le 27 novembre 2024 - 6 pages

**Document 5 :** « La qualité des données est primordiale pour libérer tout le potentiel de l'IA » - *journaldunet.com* - 23 septembre 2024 - 4 pages

**Document 6 :** « Qu'est-ce qu'une politique de gouvernance des données et comment la mettre en place » - *talend.com* - consulté le 25 février 2025 - 4 pages

**Document 7 :** « Big Data : les opportunités offertes pour les collectivités locales » - *territoireconnecte.fr* - 15 novembre 2022 - 3 pages

**Document 8 :** « Le rôle de l'IA et du machine Learning dans l'analyse de données » - *online.edhec.edu* - 28 septembre 2023 - 4 pages

**Document 9 :** « En quoi consiste l'extraction de données ? Définition, processus et outils » - *ibml.com* - consulté le 27 novembre 2024 - 4 pages

**Document 10 :** « Le top 8 des meilleurs outils de data visualisation en 2024 » - *ambient-it.net* - 11 avril 2024 - 11 pages

**Document 11 :** « Qu'est-ce que l'Internet des objets ? » - *OVH cloud* - consulté le 27 novembre 2024 - 5 pages

**Document 12 :** « Qu'est-ce que l'analyse des données ? Méthodes, techniques et outils » - *testsiteforme.com* - 19 octobre 2020 - 7 pages

**Liste des annexes :**

**Annexe A :** « Descriptif du SI de Commaglo » - 2 pages

**Documents reproduits avec l'autorisation du C.F.C.**

*Certains documents peuvent comporter des renvois à des notes ou à des documents non fournis car non indispensables à la compréhension du sujet.*

*Dans un souci environnemental, les impressions en noir et blanc sont privilégiées. Les détails non perceptibles du fait de ce choix reprographique ne sont pas nécessaires à la compréhension du sujet, et n'empêchent pas son traitement.*

# Le Modern Data Lake va-t-il se généraliser ?

La notion de Modern Data Lake commence à faire son chemin, avec pour mission d'optimiser les performances et de surmonter les défis actuels des infrastructures de données.

L'intérêt du Data Lake pour les entreprises n'est aujourd'hui plus à démontrer. Atout essentiel pour les organisations qui souhaitent pouvoir stocker, gérer et interagir rapidement avec de vastes quantités de données, son adoption est croissante. Une tendance qui est vouée à se poursuivre selon le cabinet d'étude Research and Markets, le marché des data lakes devant tripler entre 2022 et 2027, en passant de 10,2 milliards de dollars à 29,9 milliards de dollars.

Lorsque le Cloud Data Lake est apparu il y a une dizaine d'années, il a libéré les équipes data en leur permettant une gestion centralisée à distance. On voit maintenant émerger de nouveaux modèles qui ambitionnent de réinventer le système pour l'optimiser. Parmi eux, la notion de Modern Data Lake commence à faire son chemin, avec pour mission d'optimiser les performances et de surmonter les défis actuels des infrastructures, qui dérivent actuellement, dans certains cas, vers de véritables enchevêtrements de données. Face à ces prédécesseurs historiques, le Modern Data Lake incarne une avancée révolutionnaire en fusionnant performance, évolutivité, sécurité et polyvalence. Que ce soit pour l'analyse de données complexes et multi-sources ou la gestion d'informations sensibles, il peut devenir un atout majeur de la stratégie data des entreprises. Avec l'explosion actuelle des données combinée à la migration vers le cloud, ce modèle permet de répondre aux nouveaux besoins d'agilité dans le traitement des données.

## Le Modern Data Lake vs les Data Lakes historiques

Bien qu'elles aient fait leurs preuves sur plusieurs décennies, les infrastructures de données traditionnelles présentent des limites qui ne leur permettent plus de pleinement faire face aux défis actuels, tels que l'explosion de la quantité des données à traiter. Les technologies présentes au sein d'anciens Data Lakes basés sur des frameworks historiques tels qu'Hadoop ou des systèmes de stockage NAS ou noSQL, nécessitent une formation approfondie des équipes data pour les maîtriser, instaurer des règles de sécurité et un certain niveau d'évolutivité. Cette complexité d'utilisation associée à des latences rendant impossible le traitement des données en temps réel nécessitent l'utilisation de ressources humaines et financières importantes et peuvent rapidement faire de la gestion des données un véritable casse-tête.

Le Modern Data Lake va encore plus loin que les architectures de données traditionnelles grâce aux fondamentaux qui le régissent. Il permet une gestion centralisée à distance grâce à une couche d'accès fédérée permettant de se connecter à toutes les sources de données, qu'elles soient sur le cloud ou on-prem. Le Modern Data Lake intègre également une couche de sécurité et de gouvernance, de façon à pouvoir gérer l'accès, la confidentialité et le chiffrement des données, ou, pour aller plus loin, la surveillance et la journalisation. Il dispose d'un moteur de requêtes scalable et performant, accompagné de fonctionnalités propriétaires

qui optimisent le temps de réponse et une exécution tolérante aux pannes pour éviter que les requêtes de longue durée et les transformations complexes n'échouent en étant limitées par la mémoire. Et enfin une couche de modélisation et de sémantique, qui comprend toutes les fonctionnalités permettant de construire, d'organiser et de partager des modèles de données avec le langage SQL standard.

Pour en démontrer les bénéfices, il suffit d'être concret. Prenons une entreprise de services ou de retail qui souhaite analyser les habitudes de ses clients. Une architecture traditionnelle telle qu'Hadoop pourrait avoir du mal à fournir cette analyse en temps réel de par ses performances, tandis qu'une solution Modern Data Lake rend cela possible. La problématique est la même pour des développeurs qui voudraient créer une application et dont la gestion de tâches pourrait s'avérer bien plus complexe. Quant à une entreprise du secteur bancaire ou financier qui posséderait des données sensibles, les configurations avancées que nécessitent les Data Lakes historiques peuvent également représenter un défi pour garantir la sécurité des données confidentielles. À l'inverse, les Data Lakes plus modernes possèdent eux des fonctionnalités de sécurité intégrées, bénéficiant de contrôles d'accès avec des politiques de gouvernance qui en facilitent la gestion.

Un Modern Data Lake comprend de nombreux connecteurs intégrés, de façon à inclure des flux de données externes dans l'exploration des données, permettant davantage d'interopérabilité qu'avec un Data Lake traditionnel. Ce modèle offre également des fonctionnalités avancées similaires à celles d'un Data Warehouse pour permettre de gérer les opérations transactionnelles directement au niveau du Data Lake. Pour cela, une prise en charge des formats de table modernes tels qu'Iceberg, Delta Lake ou Hudi, est indispensable.

Le Modern Data Lake offre donc davantage de flexibilité, de scalabilité et d'accessibilité aux données, tout en permettant aux équipes data d'avoir recours à leurs outils habituels de data engineering, data science et business intelligence, sans en changer. Combinant les avantages du Data Lake avec les fonctionnalités d'un Data Warehouse, ce modèle est qualifié de "Data Lakehouse", offrant un environnement complet d'analyse, de gestion et de traitement de données. L'adoption d'un Data Lakehouse constitue donc la prochaine étape d'évolution pour des entreprises data-driven qui souhaitent rester compétitives et innovantes dans des secteurs fortement concurrentiels. En optimisant l'exploration de toutes leurs données, elles peuvent identifier des informations capitales pour leur développement qui ouvrent la voie à de nouvelles opportunités de croissance dans l'ère de la donnée.

## Les 7 meilleurs outils de collecte de données en 2023



Les outils de collecte de données ont changé la façon dont les études de marché sont réalisées. Ils ont transformé la collecte et l'analyse des données en un processus intelligent qui améliore les résultats de l'entreprise, ce qui signifie qu'ils sont un aspect important de toute entreprise puisqu'ils collectent, mesurent et analysent des données qui peuvent être utilisées pour les décisions de l'entreprise, la planification stratégique, la recherche et d'autres choses.

L'époque où les organisations élaboraient des stratégies sur la base d'hypothèses, d'instincts et de possibilités est révolue. Aujourd'hui, ils ne peuvent plus se permettre de commettre une erreur ou de manquer une occasion et de laisser le concurrent gagner. Les chefs d'entreprise et les analystes s'appuient sur les données pour stimuler la croissance et le chiffre d'affaires.

Les outils de collecte de données permettent de mieux comprendre les employés et les clients en recueillant leurs commentaires, leurs opinions et leurs choix. Ils peuvent contribuer à améliorer leur expérience en répondant à leurs attentes en matière de produits et de services. Pour y parvenir, il est nécessaire que les données soient suffisamment riches pour pouvoir en extraire des informations significatives. Cela permettra d'améliorer les pratiques de segmentation du marché.

Avec des données correctes, une entreprise peut déterminer comment améliorer ses performances et faire des plans pour le présent et l'avenir.

Toutefois, la collecte de données doit être effectuée de manière fiable et efficace pour que les résultats soient utilisables dans un environnement professionnel. Des données inexactes conduisent à terme à de mauvaises décisions, c'est pourquoi il est important de choisir les meilleurs outils de collecte de données.

Dans ce blog, nous verrons ce qu'est un outil de collecte de données et nous explorerons les 7 meilleurs outils de collecte de données pour vous aider à choisir le meilleur.

### Qu'est-ce qu'un outil de collecte de données ?

Les outils de collecte de données sont des instruments essentiels utilisés pour recueillir des données dans le cadre d'une recherche. Ils permettent de collecter et de stocker des données provenant de diverses sources. Ils se présentent sous différentes formes, allant de simples questionnaires papier à des applications logicielles avancées.

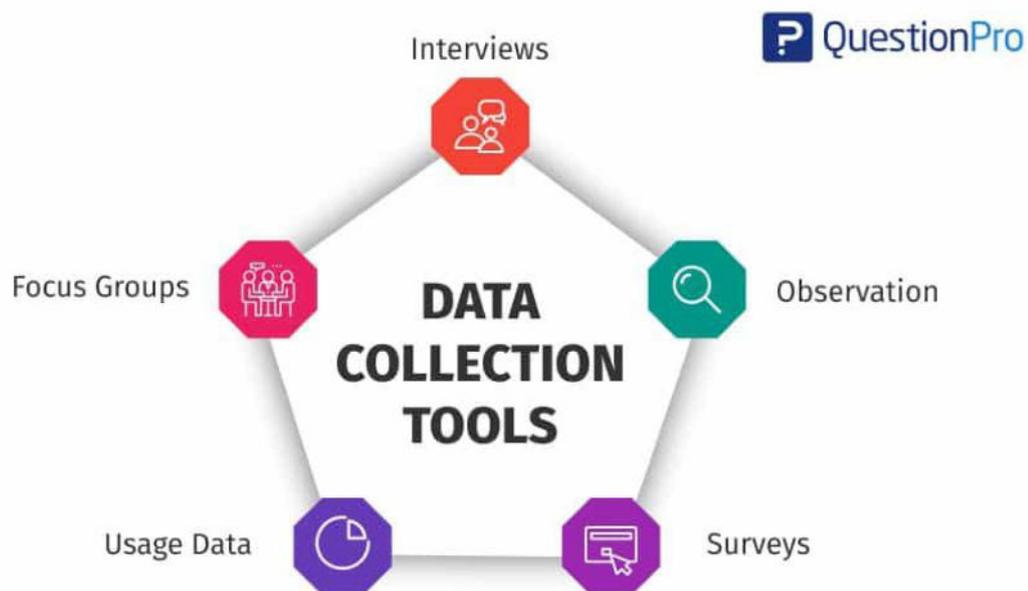
Ils peuvent être utilisés dans de nombreux domaines, comme les études de marché, l'obtention d'un retour d'information de la part des clients, le suivi du trafic sur un site web ou la collecte de données pour la recherche universitaire de manière efficace et efficiente.

Le choix du bon outil de collecte de données est crucial pour obtenir des données précises et fiables. Les chercheurs doivent tenir compte de la question de recherche, de la population étudiée et du niveau de détail souhaité lorsqu'ils choisissent l'outil de collecte de données approprié. Chaque outil ayant ses forces et ses faiblesses, il convient de choisir celui qui correspond le mieux aux objectifs de la recherche.

Il existe de nombreux types d'outils de collecte de données, et le meilleur pour une situation donnée dépend des besoins et des objectifs spécifiques de l'utilisateur en matière de données.

### Types d'outils de collecte de données

Le succès d'une étude de marché dépend de la précision des données collectées. Mais cela ne s'arrête pas là. Il est important d'analyser ces données afin que les cadres dirigeants puissent en tirer un sens.



De nombreux outils de collecte de données sont disponibles sur le marché. Le choix de la bonne solution dépend de vos besoins. Les méthodes de collecte de données les plus couramment utilisées sont les suivantes

## **Les entretiens**

Les entretiens sont une méthode de collecte de données par laquelle une personne pose une série de questions à une autre qui fait partie du public cible. L'entretien individuel peut être mené en personne ou par appel téléphonique. Les réponses sont enregistrées et analysées pour comprendre leur comportement, leurs choix et leurs expériences. Un entretien fournit des données de haute qualité car le répondant est moins susceptible de fournir des réponses fausses lorsqu'il fait face directement à quelqu'un. La limitation de ce mode de collecte de données est qu'il n'est pas réalisable pour la collecte de données à grande échelle. De plus, il est chronophage d'interviewer de nombreuses personnes sur une base individuelle. Les entretiens sont les plus appropriés lorsque le sujet est complexe ou sensible.

## **Observation**

L'observation est une technique de collecte de données dans laquelle le public cible reçoit l'énoncé du problème et ses réponses sont notées avec un minimum d'interférence de la part d'un modérateur.

Pendant que le répondant est occupé à répondre aux questions, son langage corporel, ses manières, ses expressions faciales, le ton de sa voix, etc. sont notés. Ces signes non verbaux sont importants pour la prise de décision.

L'avantage de cette méthode est qu'elle nécessite moins de préparation de la part du chercheur. L'inconvénient est que les personnes interrogées peuvent prendre conscience d'elles-mêmes si quelqu'un les observe en permanence. Cela pourrait même les mettre mal à l'aise.

L'observation est une technique de collecte de données dans laquelle le public cible reçoit l'énoncé du problème et ses réponses sont notées avec un minimum d'interférence de la part d'un modérateur. Pendant que le répondant est occupé à répondre aux questions, son langage corporel, ses manières, ses expressions faciales, le ton de sa voix, etc. sont notés.

Ces signes non verbaux sont importants pour la prise de décision. L'avantage de cette méthode est qu'elle nécessite moins de préparation de la part du chercheur. L'inconvénient est que les personnes interrogées peuvent prendre conscience d'elles-mêmes si quelqu'un les observe en permanence. Cela pourrait même les mettre mal à l'aise.

## **Enquêtes**

Les enquêtes sont une liste de questions conçues de manière à aborder les différentes caractéristiques des produits et des services. Sur la base des données collectées, divers algorithmes analytiques sont exécutés, et des rapports sont générés. Les enquêtes sont réalisées à l'aide d'outils de collecte de données en ligne qui permettent de diffuser les questions de l'enquête par courrier électronique, site web, application mobile, code QR ou plates-formes de médias sociaux .

Ces outils de collecte de données permettent de générer différents rapports tels que l'analyse des tendances, la segmentation des données, l'analyse de corrélation, l'analyse conjointe, etc. Dans les grandes entreprises, les rapports sont souvent utilisés par plusieurs équipes et doivent donc être partagés avec de nombreuses personnes .

Il est donc nécessaire d'exporter les rapports dans des formats couramment utilisés tels que .xls, .doc ou .pdf. En outre, les données de l'enquête peuvent être utilisées de manière

plus approfondie en intégrant des applications tierces à l'outil de création d'enquête. Cela multiplie l'utilité des données collectées et améliore le retour sur investissement.

Le logiciel de création d'enquêtes offre une grande flexibilité car l'utilisateur peut choisir parmi une variété de types de questions et ajouter de la logique pour personnaliser le déroulement de l'enquête en fonction des réponses données par les personnes interrogées.

Les personnes interrogées peuvent répondre à leur convenance. Si un répondant a commencé l'enquête mais ne l'a pas terminée, l'administrateur de l'enquête peut envoyer des rappels programmés et augmenter le taux de réponse. Les enquêtes permettent d'atteindre simultanément un vaste public et de collecter un grand nombre de données en un temps réduit par rapport à d'autres méthodes.

### **Données d'utilisation**

Avec l'utilisation croissante de la technologie, les données sont collectées à différents points, de la fabrication à la livraison. Ces données peuvent être utilisées pour améliorer les résultats de l'entreprise. Les enregistrements historiques, les données d'utilisation des applications, les rapports de vente, les coordonnées GPS, etc. sont autant d'exemples de données utiles qui peuvent être recueillies. Ces méthodes de collecte de données peuvent être utilisées lorsque vous souhaitez optimiser les systèmes existants au lieu d'en créer de nouveaux.

### **Groupes de discussion**

Les groupes de discussion sont constitués de personnes issues de différents domaines de la vie. Ils font l'objet d'une discussion de groupe en présence d'un modérateur. Comme les gens viennent d'horizons différents, cela conduit à explorer divers angles de pensée. On estime que la diversité des idées est source d'innovation. Les groupes de discussion aident les parties prenantes à valider les faits sur place et à recueillir un ensemble diversifié de données. Cette méthode de collecte de données souffre du risque de la présence d'une personne trop influente dans le groupe. Le modérateur doit être capable de gérer la discussion et d'équilibrer la dynamique du groupe. Les groupes de discussion sont appropriés pour les tests bêta des produits nouvellement lancés.

Les outils de collecte de données doivent être sélectionnés sur la base de faibles barrières à l'entrée, de leur capacité à s'adapter et de leurs fonctions avancées au-delà de la collecte de données de base.

### **Top 7 des meilleurs outils de collecte de données**

Il existe de nombreux types d'outils de collecte de données, tels que les enquêtes de collecte de données, les applications de collecte de données et les logiciels de collecte de données en général. Voici une liste d'outils de collecte de données que vous pouvez utiliser pour analyser les données au travail :

#### **1. QuestionPro**

QuestionPro est une plateforme d'enquête et l'un des meilleurs outils de collecte de données. Ils disposent d'un logiciel facile à utiliser qui comprend des outils pour créer, distribuer et analyser des enquêtes, des sondages, des formulaires et des questionnaires en ligne. Grâce à ces outils, les utilisateurs peuvent facilement collecter et analyser des données.

Les entreprises, les chercheurs et les organisations l'utilisent comme outil de collecte de données pour recueillir des informations et des données auprès de personnes spécifiques. C'est l'un des meilleurs outils de collecte de données, particulièrement utile pour les entreprises et les organisations qui ont besoin d'un retour d'information ou qui étudient le marché.

**Meilleure caractéristique :**

- Capacités d'analyse des données en temps réel
- Modèles préétablis
- Plusieurs types de questions et options de formatage
- Aperçu de l'enquête en temps réel
- Fonctionnalité avancée de logique de saut et de tuyauterie

**Fixation des prix**

La tarification premium de QuestionPro commence à 99,00 \$ par mois.

**2. Magpi**

Magpi est une application qui fournit des formulaires mobiles pour la collecte de données sur le terrain. Ces informations peuvent ensuite être intégrées dans des tableaux de bord et des rapports en ligne qui sont mis à jour en temps réel ou peuvent être ajoutés à des systèmes en ligne. Magpi est souvent utilisé par les entreprises qui souhaitent effectuer des enquêtes de maintenance, des inspections d'équipement, des inspections de site et des rapports d'avancement.

**Meilleure caractéristique :**

- Réponse vocale interactive
- Saisie de données hors ligne
- ramification logique

**Fixation des prix**

Le plan tarifaire de base commence à 250,00 \$, et le plan tarifaire d'entreprise commence à 100,000 \$ par mois.

**3. Retour d'information sur Zonka**

C'est l'un des outils de collecte de données et de retour d'information les plus utilisés. Zonka est également largement utilisé pour analyser les entretiens avec les clients ou les enquêtes. Vous pouvez réaliser trois types d'enquêtes : NPS, CES et CSAT.

L'expérience des clients sur le site web peut être recueillie au moyen de fenêtres contextuelles, de boutons de rétroaction, d'éléments intégrés et de liens. Grâce à l'éditeur WYSIWYG de Zonka, vous pouvez réaliser des enquêtes intéressantes en quelques minutes seulement.

**Meilleures caractéristiques :**

- Modèles préétablis

- Capacités d'analyse des données en temps réel
- Branchement conditionnel

#### **Fixation des prix**

La tarification professionnelle commence à 99,00 \$ par mois avec une possibilité d'essai gratuit.

#### **4. Zoho**

Vous pouvez créer des formulaires pour recueillir des données, les partager en ligne et recevoir des alertes rapides avec Zoho. Il existe plus de 30 types de champs différents, des thèmes modifiables, des modèles pour différentes situations et une interface utilisateur simple. Il vous permet de créer des formulaires esthétiques et utiles pour tous vos besoins.

Zoho Forms est un logiciel frontal pour les systèmes de collecte de données qui est très fiable et peut être utilisé avec de nombreuses applications différentes.

#### **Meilleures caractéristiques :**

- Autoriser le partage des liens du formulaire sur les médias sociaux
- Création de formulaires en ligne et hors ligne
- Enquêtes géolocalisées

#### **Fixation des prix**

Ils ont 4 séries de prix. La tarification commence à partir de 14,00 \$ par mois.

#### **5. Forme de papier**

Il s'agit d'un outil de collecte de données fiable qui permet à quiconque de développer facilement des formulaires ou des pages de produits. Avec Paperform, vous pouvez collecter plus de 20 types de données en ligne, notamment des fichiers, des signatures électroniques, des courriers électroniques, des textes, des adresses et des photographies.

Vos données sont également formatées automatiquement par les champs de questions, ce qui facilite l'inspection, l'exportation et l'analyse.

#### **Meilleures caractéristiques :**

- Exporter les données à tout moment au format PDF, CSV ou Word doc
- Intégration Zapier
- Les données sont stockées dans le nuage et sont protégées par SSL

#### **Fixation des prix**

Ils ont classé leurs prix en 3 catégories. Le prix de départ (essentiel) commence à 20,00 \$ par mois.

#### **6. Champ rapide**

FastField est l'un des meilleurs outils de collecte de données avec des fonctionnalités avancées qui facilitent la collecte et l'envoi de données. Il propose également des tutoriels pour vous aider à apprendre à créer des formulaires.

Cet outil de collecte de données permet également de présenter immédiatement de grandes quantités ou des types de données spécifiques sous forme de tableaux, de graphiques ou d'autres types de rapports faciles à comprendre.

**Meilleures caractéristiques :**

- Facile à utiliser
- L'intégration de Zapier est possible
- Il prend en charge plusieurs applications de formulaires mobiles

**Fixation des prix**

Le prix de FastField commence à 20,00 \$ par mois.

**7. Dispositif magique**

Device Magic est l'un des meilleurs outils de collecte de données qui vous permet de gérer vos équipes sur différents chantiers ou lieux d'enquête.

La fonction « glisser-déposer » de cet outil facilite la tâche des débutants. Ses réponses peuvent être horodatées ou géolocalisées automatiquement pour indiquer quand et où le formulaire a été rempli.

**Meilleures caractéristiques :**

- Fonctionnalité de glisser-déposer
- Horodatage automatique
- Géo estampillé

**Fixation des prix**

Contactez Device Magic pour connaître les tarifs.

**Pourquoi QuestionPro est-il l'un des meilleurs outils de collecte de données?**

QuestionPro est une plateforme de collecte de données qui vous permet de créer et de distribuer des enquêtes et de recueillir les réponses des participants. Vous pouvez utiliser la plateforme pour élaborer votre enquête, choisir votre public cible et décider de la manière dont vous souhaitez distribuer l'enquête.

Ils vous offrent de nombreuses possibilités d'analyse et de rapport sur les données collectées, telles que la visualisation des données en temps réel, l'exportation vers Excel ou CSV, et l'intégration avec d'autres outils tels que Google Analytics.

QuestionPro possède de nombreuses fonctionnalités qui peuvent vous aider à réaliser de bonnes enquêtes et à collecter de bonnes données. QuestionPro peut vous aider à collecter et à analyser des données, que vous soyez un chercheur, un chef d'entreprise ou simplement quelqu'un qui souhaite connaître l'opinion des gens. C'est pourquoi QuestionPro est l'un des meilleurs outils de collecte de données pour votre entreprise.

## Conclusion

Il est important de trouver les meilleurs outils de collecte de données lorsque votre activité dépend des données. Vous avez besoin d'une plateforme parfaite en temps réel pour collecter des données et présenter une analyse complète pour vous aider à prendre des décisions.

QuestionPro peut être la meilleure option pour vous si vous êtes à la recherche des meilleurs outils de collecte de données. Vous pouvez créer des enquêtes avec des questionnaires prêts à l'emploi et personnalisés et les distribuer par différents moyens tels que l'email et les médias sociaux ou les intégrer sur un site web.

Vous pouvez ensuite recueillir les réactions des participants sous forme de données. Enfin, vous pourrez visualiser les données collectées en temps réel et utiliser les outils d'analyse et de rapport pour approfondir les résultats.



23 Avr, 2024

Open data (données ouvertes), big data, data center... les **données** sont aujourd'hui partout et interviennent dans tous les domaines de nos vies. Les collectivités n'y échappent pas. Les données sont à la base des prises de décisions politiques sur les grands enjeux tels que la mobilité, la transition écologique et énergétique, le logement, l'aménagement, etc. Par ailleurs, la loi du 7 octobre 2016 pour une république numérique oblige les collectivités de plus de 3500 habitants et employant plus de 50 agents à partager leurs données en open data. Cela concerne environ 5000 collectivités. Ainsi, il est essentiel pour les collectivités de maîtriser la collecte, l'exploitation et la gestion des données.

### **En quoi la gestion des données est un enjeu politique et stratégique pour les collectivités territoriales ?**

Élus, direction générale, services techniques, responsables bâtiments, chacun a ses préoccupations vis-à-vis des données à son niveau. Les élus ont besoin de données pour prendre des décisions en ayant la **meilleure connaissance** possible de leur territoire. La direction générale a besoin de données pour alimenter ses tableaux de bords et fournir des éléments **d'aide à la décision**. Les services techniques et les responsables bâtiments utilisent la donnée au quotidien pour gérer leur parc de bâtiments et de réseaux, planifier la maintenance et les travaux.

Pour assurer une bonne utilisation des données par tous, il est conseillé de définir des règles et des principes de gestion. En d'autres termes, il faut élaborer une stratégie. Pour la définir, il faut néanmoins comprendre les **enjeux associés à la stratégie de gestion de données**.

### **Quels sont les enjeux associés à la stratégie de gestion de données ?**

Le premier enjeu est le respect de la **souveraineté des données**. La souveraineté désigne le fait que les données numériques doivent être soumises à la législation du pays où elles sont stockées. Pour les collectivités françaises, cela signifie par exemple que les données doivent

être conformes au Règlement Général de la Protection des Données (**RGPD**), et que les recours aux solutions des géants du web GAFAM (Google, Apple, Facebook, Amazon, Microsoft) sont à éviter.

Un enjeu majeur est l'**impact environnemental des données**. Le stockage de ces dernières est particulièrement préoccupant en raison de la forte consommation énergétique qu'il génère, constituant 19 % de la consommation énergétique totale du secteur du numérique. Une approche du « numérique responsable » doit donc être adoptée, en passant par le choix des équipements et des logiciels.

L'émergence de nouveaux usages et l'intégration de nouvelles données dans le pilotage des politiques publiques requièrent une adaptation des systèmes d'information des collectivités. L'enjeu réside dans le choix des technologies afin de garantir à la fois l'**accessibilité et la sécurité des données**. Pour répondre à ces évolutions, il est nécessaire d'intégrer de nouvelles priorités dans le développement des outils informatiques et d'ajouter des exigences dans les cahiers des charges, telles que la mise en place d'interfaces pour accéder aux données (API), l'utilisation de nouveaux modes de stockage adaptés aux données massives (comme les « lacs de données »), le choix de logiciels assurant le contrôle public de l'utilisation des données et la possibilité de les publier en open data, la capacité de recourir à l'open source, ainsi que l'exigence de formats favorisant l'interopérabilité pour faciliter la combinaison et l'exploitation optimale des données sur un territoire ou entre territoires.

Outre l'obligation de se conformer aux exigences juridiques, il y a l'enjeu de faire des choix **au service de l'intérêt général**. En effet, au-delà des obligations, persistent des questions complémentaires auxquelles il faut répondre. Quels usages prioritaires faire des données ? Quelle politique d'open data ? Quel niveau de transparence ? Comment rendre les outils accessibles à tous ? Toutes ces interrogations sont d'ordre politique, voire éthique, et constituent la doctrine des collectivités.

Aussi, les **choix contractuels** représentent un enjeu. Les collectivités doivent établir des règles claires pour contrôler et utiliser les données produites sur leur territoire, par leurs prestataires, délégataires et opérateurs. À titre d'exemple, elles rencontrent souvent des difficultés à récupérer ces données après avoir délégué des services.

Enfin, il est crucial d'**adapter l'organisation**. La mise en place d'une stratégie et d'une gouvernance de la donnée nécessite une organisation spécifique au sein des collectivités territoriales. Cette organisation peut prendre plusieurs formes selon la taille et la maturité de la collectivité. Il peut s'agir d'un(e) chargé(e) de mission data, de la désignation d'une personne en charge de la protection des données, d'identifier des référents data dans les principales directions, etc.

### **Comment mettre en place sa stratégie de gestion de données ?**

Avant toute chose, il est essentiel d'établir un **diagnostic de l'existant** pour recenser les données dont dispose la collectivité, et comment elles sont gérées. Les premières étapes après ce diagnostic sont de se **mettre en conformité** avec la loi et d'engager une démarche d'ouverture des données publiques.

La Commission nationale de l'informatique et des libertés (CNIL) recommande une approche en 6 étapes pour assurer la conformité à la gestion des données personnelles. Premièrement, chaque organisme public doit nommer un délégué à la protection des données. Ensuite, il est

nécessaire de créer un **registre de traitement** pour évaluer l'impact concret du règlement européen sur la protection des données. Sur la base de ce registre, il est essentiel d'identifier les traitements les plus sensibles et de développer un plan d'action pour renforcer leur sécurité. De plus, pour les traitements à risques élevés, il est recommandé de mener une étude d'impact (AIPD). En organisant les processus internes, les risques peuvent être anticipés et un mode de fonctionnement adapté peut être prévu en cas d'incident. Enfin, il est crucial de **suivre rigoureusement l'ensemble du processus** de mise en conformité au RGPD et de documenter la mise en œuvre des différentes étapes.

Engager une démarche d'ouverture des données publiques présuppose d'avoir formalisé la décision avec l'ensemble des services et des élus, et de former les équipes pour les impliquer et les rassurer. Au préalable, il faut aussi avoir identifié les données à ouvrir. Pour ce faire, il est possible de mettre en place une grille de critères pour **prioriser les jeux de données** (data set) à ouvrir. Ensuite, il faut procéder à l'extraction et la publication des données. Enfin, après la publication des données, il est essentiel de d'animer la démarche d'ouverture de celles-ci, pour encourager leur continuité et leur mise à jour.

Pour structurer la stratégie de gestion de données, il est essentiel de **construire une fonction data**, au même titre qu'une fonction RH par exemple. Cette fonction permettra d'établir la coordination entre tous les collaborateurs impliqués dans le cycle de vie de la donnée, de faciliter son accès, de faire vivre la stratégie data de la collectivité, tout en assurant la conformité avec les obligations légales. Comme pour la gestion d'actifs où il est indispensable de rédiger une politique de gestion d'actifs, la rédaction d'un **document de référence** pour la gestion de données est un bon début. En s'appuyant sur le RGPD, ce document fixe les règles et les objectifs de la gestion de données de la collectivité.

Parmi les réflexions à adopter à la mise en place de la stratégie de données se trouve celle sur les **choix techniques**. Ils doivent être en adéquation avec le document de référence précédemment cité.

- La première question concerne l'**hébergement des données** : faut-il préférer un logiciel en ligne (SAAS) ou un hébergement local ? en tenant compte des implications sur la sécurité, la performance et l'accessibilité des données.
- Ensuite, il s'agit de décider d'utiliser des solutions open source ou des outils fournis par des grands prestataires, en fonction des **besoins spécifiques** des données et des services concernés.
- Un autre aspect important est de trouver le bon équilibre entre transversalité et respect des verticales métier. L'architecture existante du système d'information peut poser des défis pour la circulation des données, nécessitant une réflexion sur la manière d'intégrer efficacement les données à travers les différents départements de la collectivité.
- Enfin, il est nécessaire d'examiner la pertinence de mettre en place une **plateforme de données centralisée**, permettant de rassembler toutes les données utiles grâce à des interfaces avec les systèmes d'information métier existants.

En conclusion, la gestion des données (comme nous l'avons précédemment constaté dans cet article) représente un **défi crucial pour les collectivités territoriales**, tant sur le plan politique que stratégique. La nécessité d'une stratégie de gestion des données est évidente, car elle touche à tous les niveaux de la collectivité, des élus à la direction générale en passant

par les services techniques. Une approche coordonnée et stratégique est essentielle pour garantir que les données sont utilisées de manière efficace, afin de rendre des décisions éclairées et répondre aux besoins des citoyens.

Grâce à son outil myA, tbmaestro accompagne les collectivités dans la gestion des données sur leurs parcs d'actifs. myA s'apparente en effet à un concentrateur de données, sous la forme d'une solution web (aussi appelée SAAS pour Software As A Service). Cet outil permet aux services techniques des collectivités de prendre des décisions à partir de tableaux de bords. Ils peuvent contenir l'ensemble des informations des actifs : caractéristiques (informations générales, ERP, dépenses, consommations, etc.), besoins en travaux, contrôles réglementaires, etc. L'ensemble est sécurisé et conforme au RGPD.

## DOCUMENT 4

consulté le 27 novembre 2024



La transformation numérique du secteur public place la donnée au cœur des préoccupations des collectivités territoriales. Quels sont les nouveaux enjeux et les nouvelles solutions que les collectivités peuvent prendre en compte pour offrir davantage de services numériques aux usagers et citoyens tout en gardant le contrôle sur leurs données ?

### **De l'open data au data management**

La loi pour une république numérique — aussi appelée loi Lemaire — entrée en vigueur depuis le 7 octobre 2016, s'applique désormais depuis le 7 octobre 2018 aux bases de données et aux documents d'intérêt économique, social, sanitaire ou environnemental. À compter de cette date, la plupart des administrations doivent publier gratuitement les informations rendues publiques, et dans un format à la fois ouvert et exploitable par une machine. Cette loi fait de la circulation des données la règle, et non plus l'exception !

Ce qui entraîne, en termes de flux d'échanges et de communication en ligne, des niveaux qui n'étaient pas encore concevables il y a une décennie : la donnée devient un élément de valeur ajoutée et de richesse, autant par les volumes traités que par la diversité des données accessibles.

Cette « révolution » place donc l'open data au rang des principaux leviers de transformation des politiques publiques de la donnée, tant au niveau de l'État que des collectivités. Certes, la réforme ne couvre pas la totalité des données valorisables produites par les administrations – certaines informations comme les données personnelles n'ont pas vocation à être diffusées sans contraintes [1] – mais elle devrait déjà contribuer largement au décloisonnement et à une meilleure circulation de l'information.

L'État joue un rôle moteur dans la structuration de ces politiques. Il impulse par exemple des formes d'actions publiques plus orientées sur la collaboration et l'échange de données entre administrations, avec la mise en place de programmes stratégiques (comme France Connect pour la gestion des identités numériques) ou l'ouverture de bases de données nationales et de référentiels incontournables pour les acteurs publics et privés (comme la base adresse nationale (BAN) ou le fichier SIRENE dédié aux données sur les entreprises françaises). La mission « Etalab », créée par décret le 21 février 2011, est chargée de porter cette politique ambitieuse, notamment à travers la diffusion des données publiques via un portail national.

Par ailleurs, l'État s'est également engagé dans le déploiement d'une stratégie de plates-formes afin de mettre en réseau les grandes bases de données détenues par les administrations. L'État pose ainsi les fondamentaux d'une nouvelle architecture numérique articulée par une interopérabilité fondée sur les interfaces de programmation (API). La mise en place de tels dispositifs a pour vocation de faciliter la réutilisation des données, donc de simplifier et stimuler le développement de services innovants, prenant plus fortement appui sur les approches utilisateurs.

Les grandes collectivités se sont elles aussi engagées dans cette transition. Productrices et gestionnaires de données, elles investissent dans la transformation de leurs systèmes d'information, et graduellement dans l'acquisition de nouveaux outils : systèmes de monitoring urbain, plates-formes de données ouvertes, tableaux de bords décisionnels, outils de reporting, etc. Les réformes nationales sur la fusion des régions, la création des métropoles, et la coopération intercommunale, accélèrent au sein de ces organisations la fusion des systèmes d'information et leur recentrage sur la *réutilisation*, le *contrôle* et la *diffusion* des données publiques territoriales.

### **D'une organisation en silos à une organisation fondée sur l'usage & la réutilisation des données**

Pour aider les collectivités à préparer ce projet stratégique autour de la donnée et le réaliser au mieux au sein des territoires, des solutions innovantes existent et peuvent être envisagées.

La mise en place de tels dispositifs, encore largement tributaire de la taille des entités publiques, impose certains principes pour organiser aux différentes échelles un service de la donnée reposant sur la collecte, la diffusion et l'exploitation mutualisée des données.

On note ces dernières années une évolution singulière dans la manière dont les administrations publiques et les collectivités ont commencé à gérer leurs données. L'usage de la donnée couvre, pour l'essentiel, des finalités axées sur la performance et sur la gestion de la relation citoyen. Les administrations publiques et les collectivités commencent ainsi à s'intéresser de près aux données pour optimiser leurs dépenses, lutter plus efficacement contre la fraude, prédire des risques d'incendies ou de vols de biens, pratiquer l'analyse comportementale pour mieux gérer les heures d'affluence sur l'espace public, etc. L'objectif étant de fournir des services efficaces, personnalisés et plus proches des gens.

Cependant, les applications métiers fonctionnent encore assez systématiquement en silos à partir de bases de données structurées dédiées aux métiers et généralement propriétaires. Certes, les outils décisionnels – dits de Business Intelligence – sont en mesure d'analyser des données consolidées et structurées à partir d'extractions, mais généralement dans un contexte contraint et relativement limité.

Opter pour une démarche processus devient alors un premier moyen pour piloter son organisation de manière transverse en brisant les silos. Une fois les processus modélisés, la mise en place d'indicateurs de performance et d'outils d'amélioration continue permet de rendre l'organisation plus efficiente.

En matière de protection des données personnelles, on notera ici aussi le rôle fondamental des processus dans le respect du RGPD (Règlement Général pour la Protection des Données personnelles). Pour maîtriser les problématiques de confidentialité des données, la première étape consiste à cartographier tous les processus métiers de l'organisation. Ceci permet dans un second temps d'identifier où les données personnelles sont collectées et stockées. Enfin, il est alors possible de s'assurer que tous les processus de traitement des données sont sécurisés et conformes à la norme RGPD.

D'autre part, les systèmes d'information amorcent une évolution vers des architectures orientées services et axées sur la donnée, afin d'assurer la communication entre les différentes applications du système d'information. Ces architectures se fondent sur l'usage et la réutilisation des données au-delà de leur fonction première. De fait, cette "gouvernance de la donnée" constitue un renversement de paradigme dans le mode d'organisation de l'administration, dans le sens où la donnée devient l'élément central. Les métiers sont amenés à travailler davantage en synergie, et les temps des décisions sont potentiellement raccourcis grâce à l'actualisation et à la mise à disposition permanente de l'information. [2]

### **Point de repère sur le RGPD : Le rôle fondamental des processus**

La CNIL préconise une évaluation de l'impact de la protection des données (Data Protection Impact Assessment, DPIA) pour tous les processus qui traitent des données personnelles. Les étapes d'un DPIA sont les suivantes :

- décrire le processus de traitement des données personnelles ;
- prioriser les actions à mener ;
- évaluer les risques associés et identifier comment les traiter ;
- établir comment la conformité va être documentée et démontrée.

Au-delà d'une mise en conformité au RGPD, cette politique de documentation permet de s'assurer que les processus et procédures liées à la confidentialité des données sont accessibles et compréhensibles par tous. Dans la mesure où l'impact du RGPD est important, il est fondamental d'en faire une partie prenante du Système de Management de l'Organisation plutôt que de le traiter comme un sujet ponctuel. Ceci permet également d'identifier des synergies dans les processus de l'organisation, et mieux maîtriser les coûts associés à la mise en conformité.

### **Faire converger l'approche métier & les approches transversales**

Pour réussir à mettre en place une architecture orientée services, deux approches peuvent être envisagées :

- Une approche sur un domaine particulier, le plus souvent sur un département ou un service support (données métiers, données RH, données comptables, etc.). Dans ce cas, l'accent est mis sur l'exploitation du patrimoine informationnel avec un

questionnement sur la qualité, l'usage opérationnel, voire l'enrichissement des données disponibles ;

- Une approche globale visant la cohérence transverse des données clés pour l'organisation. Il s'agit, pour l'essentiel, de données de référence massivement réutilisables (master data), vers lesquelles il conviendra de faire converger les utilisateurs. Dans ce cas, la réussite résidera dans la mise en place de processus permettant d'accéder rapidement à ces données.

Ces approches ne s'opposent pas, elles tendent plutôt à converger, du moins sur les activités de description de la donnée et de son cycle de vie. Il conviendra donc de mettre en place une organisation capable de traiter les problèmes quotidiens de la gestion des données, et d'accompagner les projets de transformation qui vont impacter le cycle de vie des données pour l'adapter au nouveau contexte. Il conviendra aussi d'utiliser des outils logiciels accélérant cette transformation, en permettant la création concertée de référentiels de données qui serviront de socle aux développements des applications métiers.

### **Vers une plate-forme unique d'intégration de données & de gestion des processus**

Les collectivités doivent donc travailler autour de leurs processus métiers et des workflows de données associés, organiser la gouvernance des données et des référentiels, et disposer d'un socle technique permettant la préparation, la qualité et la mise à disposition des données. Bien qu'il existe sur le marché de nombreux outils spécialisés pour chacune de ces attentes (outils de type BPM, outils de gestion de workflows, outils de DQM, outils de MDM...), leur assemblage reste délicat. La perspective d'adopter un système unique, dont l'usage est standardisé à l'échelle de toute l'organisation, et disposant de tous les outils fondamentaux, reste une alternative plus intéressante : plus facile à mettre en place, avec un retour sur investissement davantage assuré.

Une telle solution existe et nous souhaitons la mettre en lumière ici : il s'agit de la solution Blueway.

Blueway est une plate-forme unique d'intégration de données et de gestion des processus réunissant :

- Un socle technologique d'intégration de données, entièrement dédié au transport, à la manipulation et au contrôle de la donnée. Il expose les données via un bus applicatif intégré (*socle Data Foundation*).

Ce socle, clé de voûte de la plate-forme, permet de se connecter à différentes sources (qu'il y en ait une ou plusieurs), d'opérer des traitements sur ces données, quels que soient leurs formats (fichiers plats, web services, bases de données) et quelles que soient leurs localisations (à l'intérieur ou à l'extérieur de la collectivité). Une fois mises en cohérence, ces données peuvent être consommées en interne ou exposées à l'externe vers les partenaires de la collectivité. Trois autres usages viennent ensuite s'appuyer sur ce socle :

- Un module dédié aux référentiels de données : ce composant prend en charge la gestion de la qualité, du cycle de vie, de la consolidation, de la gouvernance, sans oublier la supervision et le monitoring des données (*module Data Gouvernance*).

- Un atelier de design de flux techniques et workflows humains à la norme BPMN. Il permet la création et la diffusion d'écrans de saisie auprès des utilisateurs, et des écrans de contrôle et de validation (*module Process Gouvernance*).
- Un module dédié à la gestion des APIs : consommation, exposition, gateway, suivi de la consommation, règles de sécurité et autorisation (*module API Gouvernance*).

Présent sur le marché de l'urbanisation du système d'information et du BPM depuis plus de dix ans, Blueway (*éponyme de la solution*) est aujourd'hui l'éditeur d'une des plates-formes d'intégration les plus puissantes et les plus complètes du marché.

Le choix de la solution Blueway repose sur trois arguments essentiels :

- Une solution complète qui à partir d'un seul environnement permet de développer des flux, mais également de les orchestrer, de gérer les erreurs, etc. De plus, la solution permet la mise en place de référentiels.
- Une interface totalement web et graphique qui facilite le développement et la maintenance des flux de données et des processus.
- Un éditeur français à taille humaine basé à Lyon avec lequel Keyrus a tissé une relation de proximité, en qualité de partenaire intégrateur de la solution.

Quelques exemples de réalisations réussies dans les collectivités

Positionné au cœur du SI, ce type de plate-forme solutionne beaucoup de problèmes rencontrés par les directions informatiques :

- Interconnexions multiples au sein du SI de la collectivité : par exemple les communications backoffice-frontoffice pour interfacier les portails citoyens avec les applications métiers ou les référentiels ; les communications backoffice-backoffice pour interconnecter les outils de gestion de courriers avec les progiciels de gestion de dossiers ; ou enfin les interconnexions avec des SI tiers des partenaires institutionnels tel que le Département pour une Mairie, la Région pour un Département, Pôle Emploi / DGFIP pour une Région, etc. Les bénéfices sont de mieux maîtriser les flux au sein du SI et faire communiquer les applications de façon synchrone ou asynchrone, mais également de centraliser, orchestrer et suivre les interactions de plus en plus complexes au sein du SI et avec l'extérieur.
- Déploiement d'un référentiel citoyen unique (RCU) pour améliorer la performance des services : lorsque plus d'une vingtaine de systèmes différents contiennent des données sur les usagers et citoyens, il est évident qu'il faut unifier ces informations pour en améliorer la qualité et l'efficacité des services proposés. Constituer un référentiel unique utilisable par tous via la mise en place d'un MDM (Master Data Management) devient alors indispensable. Le bénéfice obtenu est ici de disposer d'une vue unique et fiable des données de l'utilisateur. Du point de vue des services en interne, les utilisateurs des applications métiers sont certains d'avoir toujours les informations les plus à jour des usagers bénéficiaires. Du point de vue du système d'information, les données dupliquées liées aux usagers sont supprimées (dé-doublonnage de données) et leur centralisation renforce le continuum des services proposés par la collectivité.

Cela permet aussi au service informatique d'être plus agile en déployant plus rapidement de nouvelles applications qui se connectent au RCU via des interfaces standardisées, sans oublier la sécurité au regard du RGPD.

+++

La dématérialisation, ou plutôt la digitalisation dans le secteur public, est une tendance de fond aujourd'hui, et une véritable ambition des pouvoirs publics[3]. Il va de soi que les collectivités doivent offrir davantage de services numériques aux usagers et citoyens tout en gardant le contrôle sur leurs données.

---

[1] Dispositions applicables issues du RGPD et de la loi du 20 juin 2018 sur la protection des données personnelles. [2] Collecte et gestion des données numériques pour le pilotage des politiques publiques, Fédération nationale des collectivités concédantes et régies (FNCCR), novembre 2016. [3] <https://www.gouvernement.fr/action/action-publique-2022-pour-une-transformation-du-service-public>.

# La qualité des données est primordiale pour libérer tout le potentiel de l'IA

La qualité des données est centrale pour une IA fiable, impartiale et efficace. Sans données précises et cohérentes, même les systèmes d'IA les plus sophistiqués peuvent générer des résultats erronés.



Image par torstensimon de Pixabay

À mesure que l'intelligence artificielle (IA) et les algorithmes d'apprentissage automatique sont de plus en plus intégrés dans divers secteurs, leur impact sur les processus décisionnels ne cesse de croître. Que ce soit pour des diagnostics médicaux, des prévisions financières ou des services clients, ces systèmes reposent sur des volumes de données croissants pour fonctionner efficacement. Cependant, l'efficacité de ces systèmes est étroitement liée à la qualité des données utilisées pour leur entraînement. Des données de mauvaise qualité peuvent mener à des prédictions inexactes, à des résultats biaisés, voire à des conséquences éthiques néfastes, compromettant ainsi l'objectif même de l'IA.

En effet, la qualité des données est essentielle pour l'IA et les algorithmes, et joue un rôle majeur pour garantir l'exactitude, l'équité, la transparence et la conformité. À l'aide d'exemples concrets, il est important de mettre en évidence les dangers de l'utilisation de données erronées et l'importance d'investir dans des ensembles de données fiables et de haute qualité pour construire des systèmes d'IA performants. La qualité des données a un impact direct sur l'exactitude, la fiabilité et les implications éthiques de l'IA et des algorithmes.

Le coût du nettoyage et du prétraitement des données augmente lorsque la qualité des données est faible. Les modèles d'IA passent une partie importante de leur cycle de développement dans la phase de préparation des données, et plus la qualité des données est mauvaise, plus il faut de ressources pour les rendre utilisables. Cela peut entraîner des coûts opérationnels plus élevés et des retards dans le déploiement des modèles d'IA.

Passons en revue les points clé et les exemples qui démontrent pourquoi la qualité des données est essentielle pour des systèmes d'IA efficaces :

# 1. Exactitude dans la prise de décision stratégiques ou opérationnelles

L'une des raisons les plus critiques d'une qualité de données élevée est la nécessité de prédictions et de décisions précises. Les algorithmes d'IA et de machine Learning apprennent à partir de données historiques pour faire des prédictions futures. Si les données sont incomplètes, incohérentes ou incorrectes, le modèle peut apprendre des modèles erronés et produire des résultats inexacts.

Prenons l'exemple d'un modèle d'apprentissage automatique conçu pour prédire les résultats des patients dans le domaine de la santé. Si les données d'entraînement contiennent des dossiers médicaux manquants ou incorrects, le modèle pourrait mal classer les conditions des patients, ce qui entraînerait des conséquences potentiellement mortelles ou grandement handicapantes telles que des diagnostics ou des traitements incorrects. Des données médicales validées de haute qualité garantissent que les systèmes d'IA font des prédictions précises, préservant ainsi le bien-être des patients.

## 2. Biais et équité dans les modèles d'IA

Le biais est un problème important qui découle de la mauvaise qualité des données. Si les données utilisées pour entraîner un système d'IA sont biaisées, le modèle d'IA héritera de ces biais et les exacerbera éventuellement. Cela a une incidence majeure sur le bon fonctionnement de vos modèles d'IA et peut conduire à des résultats discriminatoires, en particulier dans des domaines sensibles comme la justice pénale, l'embauche ou les prêts.

Si nous pensons entraîner des chatbots sur la base de conversations téléphoniques enregistrées dans les années 1960 aux États-Unis, nous aurons des résultats racistes comme le confirmait Luc Julia, co-inventeur de Siri. Cela s'est déjà produit en 2016 avec Microsoft qui a retiré « Tay » de réseau social X (ex Twitter) après avoir écrit des commentaires racistes et misogynes, un programme d'intelligence artificielle, conçu pour interagir avec les internautes. Et en 2018 c'est Amazon qui désactive une IA qui discriminait les candidatures de femmes à l'embauche. Cet algorithme a été formé sur des CV soumis sur une longue période pour l'entraîner, mais la plupart des candidats retenus étaient des hommes, ce qui a pénalisé les CV avec des mots-clés spécifiques aux femmes. Il s'agit d'un cas clair de déséquilibre des données et de biais historique, où la mauvaise qualité des données ne représentant qu'un seul groupe démographique ayant conduit à des décisions d'embauche biaisées. En veillant à ce que les données soient représentatives et équitables, corriger les bases, les compléter et faire des tests réguliers pour ne mise en qualité avant usage, ces biais pourraient être d'abord minimisés et aider à disparaître dans la durée. D'où la vigilance demandée à ne pas laisser la machine décider sans intervention humaine au début pour interpréter les résultats.

## 3. Principe GIGO (Garbage In, Garbage Out)

Le célèbre principe du « garbage in, garbage out » s'applique directement à l'IA. Si les données d'entrée sont erronées, même le modèle d'IA le plus avancé produira des sorties défectueuses. Des données de haute qualité permettent au modèle de reconnaître des modèles pertinents et de générer des informations fiables, tandis que des données de mauvaise qualité conduisent à des résultats peu fiables et souvent inutiles.

Sur les marchés financiers, les algorithmes de trading s'appuient fortement sur des flux de données en temps réel pour prendre des décisions d'achat et de vente. Si ces flux de données

contiennent des informations incorrectes ou retardées, l'algorithme peut effectuer de mauvaises transactions, ce qui entraîne des pertes financières substantielles. D'autre part, des données propres et précises permettent d'élaborer des stratégies de trading optimales et en minimisant les risques.

Dans la conduite autonome, les systèmes avancés d'aide à la conduite (ADAS – Advanced Driver Assistance Systems) sont basés sur les mêmes principes que le ChatGPT et d'autres grands modèles de langage (LLM- Large Language Models). Ils sont formés sur des images et des données de capteurs provenant de plusieurs environnements comprennent des images qui peuvent être de mauvaise qualité (par exemple, des données floues, mal étiquetées ou incomplètes), le modèle peut ne pas reconnaître les dangers routiers dans des environnements inconnus. Des données d'entraînement diversifiées et de haute qualité permettent à l'algorithme de mieux se généraliser à de nouvelles situations, améliorant ainsi la sécurité et la fiabilité.

## **4. Transparence et explicabilité**

Dans de nombreuses applications, en particulier dans les secteurs réglementés tels que la santé, la finance ou l'assurance, les systèmes d'IA doivent être transparents et explicables. Cela signifie que les décisions prises par ces systèmes doivent être compréhensibles par les humains et des rétro-ingénierie complètes. Des données de mauvaise qualité rendent difficile cette tâche, car les décisions de l'IA peuvent être influencées par des valeurs aberrantes ou des erreurs, ce qui rend plus difficile l'explication de son comportement.

IBM dans une étude publiée en 2024 estime que les erreurs de données ou aux problèmes d'explication nécessitant des corrections peuvent coûter jusqu'à 3,1 trillions de dollars par an aux entreprises retardant ainsi le déploiement de projets d'IA et entraînant des surcoûts.

Dans le secteur de l'assurance, les entreprises utilisent des modèles d'IA pour calculer les primes des clients. Si les données consommées par les moteurs décisionnels d'IA sont aberrantes, ou avec des valeurs manquantes ou incorrectes, il devient difficile pour les compagnies d'assurance de justifier leurs décisions tarifaires. Aussi, l'analyse de données et leur comparaison avec des modèles de fraude augmentés par l'IA peut alerter sur des comportements douteux de certains clients examinant des milliers d'informations pour dégager des schémas inhabituels ou des incohérences pour détecter et faire remonter des fraudes potentielles, si ces blocages sont inopérants ou injustifiés, les pertes financières peuvent être importantes. Des données de haute qualité garantissent la transparence des modèles d'IA et la défense de leurs décisions en cas d'audits ou de litiges sont de plus en plus demandées par les instances de régulation et de contrôle.

## **5. Impératifs éthiques et impact social**

Les implications éthiques de l'IA sont directement et principalement liées à la qualité des données qu'elle traite. Une mauvaise qualité des données peut entraîner des résultats biaisés, injustes ou nuisibles, érodant la confiance dans les systèmes d'IA. Il est essentiel d'assurer une qualité élevée des données pour créer des systèmes d'IA éthiquement solides et socialement responsables.

La technologie de reconnaissance faciale pilotée par l'IA a été critiquée pour avoir mal identifié des individus, en particulier parmi les groupes minoritaires. Cela découle du fait que de nombreux algorithmes de reconnaissance faciale sont entraînés sur des ensembles de données avec une surreprésentation de certains groupes ethniques et une sous-représentation d'autres. Il s'agit d'un cas clair où une mauvaise qualité des données conduit à

des résultats contraires à l'éthique et discriminatoires, et l'amélioration de la qualité des données est la clé pour résoudre ces problèmes.

La qualité des données est la base sur laquelle reposent une IA et des algorithmes efficaces. Qu'il s'agisse de faire des prédictions précises, de réduire les biais, d'assurer l'équité ou de maintenir la conformité légale, la qualité des données joue un rôle central dans la détermination du succès des systèmes d'IA. Des données de mauvaise qualité entraînent des résultats peu fiables, biaisés et contraires à l'éthique, tandis que des données de haute qualité conduisent à des solutions d'IA plus précises, transparentes et socialement responsables.

Garantir la qualité des données nécessite un investissement initial important dans les processus de collecte, de nettoyage et de validation des données, mais les avantages en termes de meilleures performances de l'IA, d'économies de coûts et d'atténuation des risques en valent la peine. En se concentrant sur des données de haute qualité, les organisations peuvent améliorer la précision de leurs modèles, minimiser les biais, maintenir la conformité aux réglementations et favoriser la confiance avec les utilisateurs. Cela permet non seulement d'améliorer l'efficacité opérationnelle, mais aussi d'aider les entreprises à créer des systèmes d'IA évolutifs, éthiques et adaptables aux défis du monde réel.

En fin de compte, la qualité de l'IA dépend des données dont elle s'inspire. En privilégiant la qualité des données, les organisations peuvent libérer tout le potentiel de l'IA, en prenant des décisions plus intelligentes, plus justes et plus efficaces qui ont un impact positif sur les activités des entreprises et sur la société dans son ensemble. Des données de haute qualité ne sont pas seulement une nécessité technique, c'est un atout stratégique dans la perspective d'usage de l'IA plus massif.

# Qu'est-ce qu'une politique de gouvernance des données et comment la mettre en place ?

Dans le monde des entreprises actuel, rapide et hautement compétitif, la gouvernance des données est indispensable. Les organisations peuvent aujourd'hui collecter d'immenses quantités de données internes et externes, mais elles ont besoin d'une discipline pour augmenter la valeur apportée par ces données, gérer les risques et réduire les coûts.

## Qu'est-ce que la gouvernance des données ?

**La gouvernance des données est un ensemble de processus, rôles, règles, normes et métriques permettant d'assurer une utilisation efficace et efficiente des informations, dans le but d'aider les entreprises à atteindre leurs objectifs.** Elle définit les procédures et les responsabilités garantissant la qualité et la sécurité des données au sein d'une entreprise ou d'une organisation. Elle définit également qui peut effectuer quelle action, sur quelles données, dans quelle situation et selon quelle méthode.

Une stratégie de gouvernance des données claire est fondamentale pour toute organisation traitant les big data, et explique comment la société peut bénéficier de procédures et de responsabilités communes et cohérentes. Les moteurs opérationnels déterminent quelles données doivent être soigneusement contrôlées dans votre stratégie de gouvernance des données ainsi que les bénéfices attendus. Cette stratégie sera la base de votre cadre de gouvernance des données.

Par exemple, si l'un des moteurs opérationnels de votre stratégie de gouvernance des données est d'assurer la confidentialité des données de santé, toutes les données relatives à des patients devront être gérées de manière sécurisée lors de leur traitement au sein de votre entreprise. Les exigences en matière de conservation (c'est-à-dire l'historique des changements apportés aux informations, avec leur date et leur auteur) seront définies pour assurer la conformité aux exigences réglementaires applicables, telles que le RGPD.

La gouvernance des données garantit que les rôles relatifs aux données sont clairement définis, et que les responsabilités sont comprises et acceptées par tous dans l'entreprise. Un cadre de gouvernance des données bien conçu couvre les rôles et responsabilités stratégiques, tactiques et opérationnels.

## Ce que n'est pas la gouvernance des données

La gouvernance des données est souvent confondue avec d'autres termes et concepts relativement proches, dont la *gestion des données* et la *gestion des données de référence* (*Master Data Management, ou MDM*).

### La gouvernance des données n'est pas la gestion des données

La gestion des données est la gestion de l'ensemble des besoins en données d'une organisation. La gouvernance des données est le composant principal de la gestion des données, celui qui relie les neuf autres disciplines, dont la qualité des données, la gestion des

données de référence (MDM), la sécurité des données, les opérations de bases de données, la gestion des métadonnées et le data warehousing.

## La gouvernance des données n'est pas la gestion des données de référence

La gestion des données de référence (MDM) se concentre sur l'identification des entités clés d'une entreprise, puis sur l'amélioration de la qualité de ces données. Elle aide à assurer l'exactitude et l'exhaustivité des informations sur des entités clés telles que les clients, fournisseurs, prestataires de soins, etc. Ces entités étant partagées par toute l'organisation, la gestion des données de référence vise à regrouper les vues partielles de ces entités dans une vue d'ensemble. Cette discipline va au-delà de la gouvernance des données.

Cependant, il ne peut y avoir de gestion des données de référence sans gouvernance correcte. Par exemple, un plan de gouvernance des données définit les principaux modèles de données (définition d'un client, d'un produit, etc.), les règles de conservation de données ainsi que les rôles et responsabilités pour l'écriture de données, l'agrégation de données et l'accès.

## La gouvernance des données n'est pas data stewardship

La gouvernance des données permet d'assigner les responsabilités appropriées en matière de données à chaque personne. Le data stewardship permet, lui, de s'assurer que les données sont exactes, contrôlées, et faciles à découvrir et à traiter par les intervenants correspondants. La gouvernance des données intervient plutôt au niveau de la stratégie, des rôles, de l'organisation et des règles, alors que le data stewardship se focalise sur l'exécution et l'opérationnalisation.

Les arbitres de données gèrent les éléments de données et s'assurent que les données elles-mêmes sont conformes au plan de gouvernance des données, liées à d'autres éléments de données et maîtrisées en termes de qualité, de conformité et de sécurité.

## Principaux avantages de la gouvernance des données

Une stratégie de gouvernance des données efficace offre de nombreux avantages à une organisation, notamment :

- **Une compréhension commune des données** : la gouvernance des données offre une vue cohérente des données et une terminologie commune, tout en laissant aux unités opérationnelles la flexibilité dont elles ont besoin.
- **Une meilleure qualité des données** : la gouvernance des données crée un plan assurant l'exactitude des données, leur exhaustivité et leur cohérence.
- **Une carte des données** : la gouvernance des données offre une fonctionnalité avancée de compréhension de l'emplacement de toutes les données liées aux entités clés, facteur indispensable de l'intégration des données. Tout comme un GPS permet de représenter une zone géographique et de guider les utilisateurs, la gouvernance des données rend les éléments de données plus faciles à utiliser et à connecter avec des résultats commerciaux.
- **Une vue à 360 degrés de chaque client et des autres entités de l'entreprise** : la gouvernance des données établit un cadre dans lequel l'organisation partage « une seule version de la vérité » et assure la cohérence nécessaire entre toutes les entités et activités de l'entreprise.

- **Une conformité à tous les niveaux** : la gouvernance des données offre une plateforme permettant de répondre aux exigences réglementaires, telles que le RGPD (Règlement général sur la protection des données) de l'Union européenne ou la loi HIPAA (Health Insurance Portability and Accountability Act) des États-Unis, et aux autres exigences du secteur, telles que les normes PCI DSS (Payment Card Industry Data Security Standards).
- **Une meilleure gestion des données** : la gouvernance des données apporte une dimension humaine dans ce monde des données fortement automatisé. Elle établit des codes de conduite et des bonnes pratiques en matière de gestion des données, afin de répondre aux besoins et inquiétudes dans tous les domaines : données et technologies bien sûr, mais aussi sécurité, conformité et autres aspects juridiques.

## Gouvernance des données dans le cloud

Puisque de plus en plus de sociétés et d'organisations prennent conscience des bénéfices que peut engendrer la migration de tout ou partie des stockages et traitements de données vers des stratégies d'intégration cloud et des solutions iPaaS, le besoin d'une gouvernance des données efficace augmente fortement.

Pour évoluer vers le cloud, il faut déléguer certaines tâches à des tiers, par exemple la gestion de l'infrastructure, le développement d'applications, la sécurité, etc. Le cloud implique également de virtualiser les ressources techniques, ce qui peut être problématique au regard de la souveraineté des données (certaines réglementations imposent par exemple que les données résident dans un lieu ou un pays précis). De plus, les stratégies privilégiant le cloud encouragent généralement la décentralisation, ce qui peut permettre à certaines branches de l'entreprise ou certains groupes de travail de créer leurs propres systèmes indépendants, et ainsi entraîner une dispersion des données non contrôlée.

La gouvernance permet de répondre à ces problématiques. Tout d'abord, un plan de gouvernance des données stratégique est indispensable pour toute migration de données vers le cloud. Que l'organisation choisisse un modèle hybride ou entièrement basé sur le cloud, le processus de migration des données sera toujours amélioré par un plan de gouvernance des données global, et la migration en elle-même sera plus efficace et plus sûre.

En outre, le déplacement des processus de traitement des données vers le cloud ajoute une couche de complexité en matière de sécurité et d'accès. Certes, les solutions de données sur site ont aussi besoin d'une stratégie de gouvernance des données solide, mais les responsables apprécient particulièrement la valeur de la gouvernance des données lorsque les données migrent vers le cloud.

## Outils de gouvernance des données

Afin d'adopter la bonne approche de gouvernance pour votre organisation, optez pour des outils open source évolutifs, qui peuvent être intégrés rapidement et à faible coût avec votre environnement.

De plus, une plateforme cloud permet une connexion rapide à des fonctionnalités robustes, faciles à utiliser et abordables. Les solutions cloud évitent également les frais supplémentaires liés aux serveurs sur site.

Lorsque vous comparez les outils de gouvernance des données, retenez ceux qui vous aideront à obtenir les avantages commerciaux que vous avez définis dans votre stratégie de gouvernance.

Ces outils doivent vous aider à :

- **Collecter et comprendre vos données**, grâce à des outils et des fonctionnalités de découverte, de profilage et de comparaison. Par exemple, les outils performants peuvent automatiquement détecter une donnée personnelle, comme un numéro de sécurité sociale, dans un nouveau dataset et déclencher une alerte.
- **Améliorer la qualité des données** avec la validation, le nettoyage et l'enrichissement des données.
- **Gérer vos données** grâce aux processus ETL et ELT basés sur les métadonnées, et aux applications d'intégration de données, afin que les pipelines de données puissent être suivis avec un historique des données de bout en bout.
- **Contrôler vos données** avec des outils de vérification et de surveillance actives.
- **Documenter vos données** afin que des métadonnées puissent leur être ajoutées, améliorant ainsi la pertinence, la recherche, l'accès, les liaisons et la conformité.
- **Responsabiliser les personnes qui connaissent le mieux les données**, en leur permettant de contribuer aux tâches de data stewardship avec des outils en libre-service.

## La gouvernance des données est une obligation

Les organisations disposent aujourd'hui d'une quantité de données incroyable sur les clients, les fournisseurs, les patients, les employés, etc. Si une organisation utilise correctement ces informations pour mieux comprendre son marché et son audience cible, elle accroît ses chances de succès. Cette gouvernance des données garantit également que les données sont fiables et bien documentées, que la recherche et l'accès à ces données sont faciles au sein de l'organisation, et que les données restent sécurisées, conformes et confidentielles.

Assurez-vous que votre organisation rentabilise au mieux ses investissements en matière de gouvernance des données et minimise les risques de violation des données. Lorsque vous êtes prêt à vous lancer, découvrez nos solutions de gouvernance des données.

## DOCUMENT 7

territoireconnecte.fr - 15 novembre 2022

# « Big Data : les opportunités offertes pour les collectivités locales »

*Les collectivités locales sont chargées de servir l'intérêt général sur leur territoire mais la multiplicité des secteurs d'intervention rend la tâche toutefois difficile. Il faut d'abord cerner les attentes de la population, puis analyser les données disponibles et enfin agir en conséquence. Il existe plusieurs solutions interdépendantes à l'égard de ces préoccupations, comme l'utilisation de capteurs et d'outils de surveillance, recueillir et traiter les données, et prendre les mesures adaptées. Ces actions renvoient à la notion de Big Data. Explications.*

### **Big Data : définition et contexte réglementaire**

**Le Big Data désigne une grande quantité de données stockées dans un seul volume.** Il tire sa performance de la complexité de son fonctionnement. Les informations à traiter proviennent de différentes sources numériques et permettent de gérer divers outils simultanément. Cependant, cet important volume de données n'altère en rien la rapidité des transferts. **Au contraire, le Big Data relaie les informations à une grande vitesse.**

Il est alors possible d'assurer une **surveillance en temps réel**. Il est également conçu pour **s'adapter à la croissance des données** à exploiter. Google et Yahoo font partie des pionniers dans l'utilisation de cette technologie.

Le Big Data joue un rôle important dans la société moderne. Les collectivités territoriales récoltent les données numériques afin de **mieux gérer les services publics** et leur fonctionnement interne.

Le cloud computing, les médias sociaux, les téléservices, les retransmissions vidéo et l'open data sont les principales sources d'informations que le Big Data recueille **dans le respect du RGPD**. Celui-ci régit la collecte des données et impose des contraintes réglementaires aux entreprises européennes. Son rôle est essentiellement de **filtrer et de protéger les données personnelles**. Le RGPD autorise d'ailleurs la Commission nationale de l'informatique et des libertés (CNIL) à sanctionner les entreprises qui dérogent à ses règles avec des amendes allant jusqu'à 20 millions d'euros ou 4 % de la valeur du chiffre d'affaires.

Les collectivités territoriales ont la charge de la **gestion et du contrôle des informations** surtout en matière de téléservices. Elles avertissent ainsi la CNIL et les victimes en cas de non-respect des données relatives à telle ou telle personne.

Les collectivités sont d'ailleurs tenues par certaines mesures dictées par le RGPD comme ne pas nuire à la liberté d'action des utilisateurs ni à leur vie privée. Elles doivent bien encadrer les sous-traitants, ainsi que détecter et gérer les accidents de traitement dès leur survenance afin de garantir la stricte confidentialité des données personnelles.

## **Big Data : d'où proviennent les données ?**

Saviez-vous que nous créons environ 2,5 milliards de milliards d'octets de données chaque jour ? Ces données ne sont exploitables que par des Big Data qui récoltent les **informations provenant d'objets connectés à Internet**.

Les données sont recueillies depuis les publications ainsi que les échanges téléphoniques et électroniques. Elles arrivent dans un centre de traitement sous leurs formes d'origine : **image, texte, vidéo et son**. Le Big Data traite aussi les informations provenant de différents capteurs : météo, taux de pollution, indication du GPS, etc. Il accède en outre aux **données dématérialisées mises en archivage**.

Les citoyens ont besoin d'accéder à diverses informations et en temps réel. Pour satisfaire cette demande, des collectivités ont dû créer des applications dédiées. Les données accessibles sont regroupées sur une plateforme locale sous forme de tableau de bord. Les informations sont automatiquement et pratiquement mises à jour instantanément.

La notion de **Data mining** est née avec le Big Data. C'est le processus permettant de **collecter, d'analyser et de retransmettre les données numériques sous une forme facilement exploitable**. Cette fonctionnalité innovante est intéressante aussi bien sur le plan marketing que dans le quotidien des citoyens. Ces derniers accèdent ainsi à des **informations structurées et en temps réel, le tout en préservant leur anonymat**.

La technologie du Big Data est exploitée par de nombreuses entreprises comme Orange Application for Business. Son algorithme Flux Vision permet par exemple d'**analyser le flux de déplacement en ville et le niveau de fréquentation des lieux**. Le suivi par le Wi-Fi est possible seulement lorsqu'une zone déterminée est couverte par le réseau public.

La vidéosurveillance est un moyen de **surveiller les zones sensibles** dans une localité donnée. Cette solution est pratique en cas de crise, mais aussi pour contrôler le fonctionnement des services et assurer la sécurité interne. Les caméras sont généralement placées au niveau des postes de commandement, des centres hydrauliques, etc.

## **Exploitation des données : cas d'usage pour les collectivités locales**

Les collectivités locales utilisent les données du Big Data pour améliorer la gestion et la prestation des services publics. Il s'agit d'abord de **contrôler à distance les besoins en énergie** de la communauté afin de prendre les mesures adaptées. Des outils peuvent par exemple être mis en place pour surveiller la consommation locale d'électricité et de gaz. Des capteurs installés à des endroits stratégiques permettent d'ailleurs de relayer les informations nécessaires via des outils à faible consommation.

Le projet Bento, expérimenté dans la métropole de Rennes, est conçu pour **établir la cartographie énergétique des bâtiments**. Une application a également été développée afin de suivre et exploiter les informations fournies.

La sécurité des citoyens est aussi au cœur des préoccupations des collectivités locales. Les mesures à prendre commencent par l'**installation d'outils de surveillance des locaux**. À titre d'exemple, New York a mis au point un algorithme analysant 60 facteurs d'incendie. Il est également du ressort de la collectivité d'identifier et de surveiller les quartiers à haut risque de délinquance.

La technologie actuelle permet également d'analyser les facteurs déclencheurs de telle ou telle maladie afin d'adopter des mesures préventives. Les cas des patients sont également regroupés dans le but de trouver rapidement les remèdes.

Le problème de la circulation est en outre un casse-tête pour les grandes villes. L'enjeu est de **fluidifier le trafic urbain à chaque moment de la journée** et de dénouer les goulots. Le projet Optymod, expérimenté à Lyon, prévoit le trafic grâce aux données fournies par les différents moyens de transport.

Dans la métropole de Rennes, c'est le projet BigCycleData qui apporte la solution. Un système de **bicyclettes en libre-service** a été mis en place pour remédier aux pics de trafic urbain. 3DEXPERIENCE City est la plateforme chargée de localiser les vélos et d'établir les habitudes des utilisateurs.

C'est en analysant les problèmes de circulation qu'il est possible de trouver la solution adaptée comme réaménager les voies, prioriser le passage des piétons aux carrefours, etc.

Une ville moderne se doit aussi de **maîtriser les paramètres environnementaux** dans sa région. Il s'agit à la fois d'analyser la qualité de l'air, les caractéristiques de l'eau et l'état des sols.

**L'eau doit être propre et gérée efficacement.** La collectivité veille à une distribution équitable de la ressource et limite les fuites. Parallèlement, la pollution entraîne des problèmes de santé pour les citoyens. Il est donc impératif de **contrôler la qualité des sols et de l'air via des applications spécifiques**. D'un autre côté, certaines expérimentations permettent de prévenir les inondations ou d'analyser les activités sismiques d'une région.

En somme, l'avancée technologie actuelle joue en faveur des collectivités locales et des citoyens. Les outils déjà mis en place ou expérimentés sont conçus pour **faciliter la gestion des activités** de chacun. Dans les **villes connectées**, rien n'est laissé au hasard pour améliorer les conditions de vie. Les citoyens ont accès aux informations essentielles en quelques clics.

De leur côté, les dirigeants des collectivités gèrent mieux leur territoire grâce aux algorithmes de surveillance. Les Big Data détiennent un rôle important dans ces processus. Ils permettent de synthétiser les données numériques recueillies pour les traduire en informations simples.

# Le rôle de l'IA et du machine learning dans l'analyse de données

Le développement technologique est un **cercle vertueux**. L'émergence d'une nouvelle technologie et son développement impliquent de créer de nouvelles solutions pour en optimiser l'impact ou les bénéfices. Et cela vaut tout particulièrement pour la data. À mesure que la quantité et la complexité des données croît, notre capacité (humaine) à en tirer parti semble décroître. D'où **l'importance de s'appuyer sur l'analyse de données pour faciliter leur exploitation**.

Parmi les outils qui sont les plus à même d'améliorer la pratique de l'analyse de données (et donc de simplifier le travail des Data Analysts), on retrouve l'intelligence artificielle (IA) et le machine learning (ML). Deux technologies de plus en plus prédominantes dans la sphère professionnelle en général et l'écosystème de la data en particulier.

Elles sont d'autant plus symptomatiques de ce fameux cercle vertueux qu'elles sont elles-mêmes alimentées par de grands volumes de données. Mais **en quoi peuvent-elles réellement simplifier ou optimiser l'analyse de la data** ? C'est le sujet que nous allons creuser dans cet article...

## Qu'est-ce que l'analyse de données ?

L'analyse de données est le processus qui consiste à **étudier des volumes plus ou moins importants de données** et à en tirer des modèles analytiques. Ce domaine de la Data Science comprend à la fois :

- **l'analyse descriptive** : soit le fait d'analyser les données et de décrire ce qu'elles racontent à un instant T ;
- **l'analyse prédictive** : prédire un potentiel résultat à partir des données que l'on a extraites d'activités passées ou actuelles.

**La collecte de données est de moins en moins un défi** pour les entreprises. Les canaux par lesquels les organisations peuvent créer et récupérer de la data sur leur activité ou leur public cible (pour ne citer que ces deux domaines d'application de la Data Science) sont de plus en plus divers et accessibles.

Néanmoins, l'analyse de volumes de données toujours plus importants est **la nouvelle bête noire des entreprises**. En effet, le Big Data n'est d'aucune utilité s'il n'est pas traduit en des termes compréhensibles par tous les collaborateurs. Et surtout, si les données ne sont pas utilisables sous la forme d'une stratégie commerciale.

**Les défis actuels posés par l'analyse du Big Data sont nombreux**. De la rationalisation de données dispersées à leur conversion en modèles prédictifs, puis en décisions opérationnelles, les Data Analysts ont du pain sur la planche... Heureusement, l'IA et le machine learning ont le potentiel de faire avancer les choses à toute vitesse !

## En quoi l'IA peut-elle faciliter l'analyse de données ?

Comme on l'a déjà observé dans des secteurs comme le marketing ou même les ressources humaines, **l'intelligence artificielle a provoqué un véritable bouleversement** dans la manière dont de nombreux secteurs opèrent. Celui de l'analyse de données (et du Big Data en général) ne fait pas exception.

La raison de ce potentiel transformateur de l'IA est sa capacité à **accélérer n'importe quel processus de traitement ou d'analyse**. Prenons par exemple le secteur du marketing. L'intelligence artificielle offre aux marketeurs la possibilité d'exploiter de grands volumes de données, afin de personnaliser leur contenu. Ou encore, de mieux prédire le comportement et les attentes des consommateurs.

L'IA et l'analyse de données sont d'autant plus intimement liées que **l'intelligence artificielle nécessite de grandes quantités de données** pour s'entraîner. C'est-à-dire, pour apprendre à réfléchir et prédire des résultats comme le ferait une intelligence humaine.

Dans le domaine de l'analyse de données, **l'objectif de l'IA est de mâcher le travail aux humains**. En effet, lorsqu'elles sont réalisées par un être humain, la collecte, le tri ou encore l'uniformisation de données venant de sources disparates nécessitent un travail considérable.

L'IA supprime en grande partie cette contrainte, car elle **peut effectuer la grande majorité de ces tâches de manière autonome**. Mais aussi, beaucoup plus rapidement...

## Quelles sont les différences entre l'analyse de données s'appuyant sur l'IA et l'analyse des données traditionnelle ?

Les principales différences entre l'analyse de données facilitée par l'IA et l'analyse de données traditionnelle s'expriment ainsi en termes de **capacités et d'autonomie informatique**. Le cerveau humain a en effet des limites en termes de quantité de données qu'il peut traiter efficacement ! Il est possible de pousser ses limites avec une formation et beaucoup de pratique, mais seulement jusqu'à un certain point. Et généralement, pour un ratio investissement/rentabilité peu intéressant pour les entreprises.

Les systèmes de data analyse s'appuyant sur l'IA peuvent **traiter des volumes de données beaucoup plus importants**. Ils peuvent également obtenir des résultats bien plus précis et réduire le risque d'erreurs.

## L'IA explicable (ou XAI)

De manière générale, c'est **en matière d'analyse prédictive des données** que l'IA se révèle la plus intéressante. L'intelligence artificielle et le Big Data vont en effet combiner leurs forces pour aider les entreprises à utiliser la data de manière plus précise et pertinente. Grâce à leurs prédictions, alimentées par de larges volumes de données, les organisations peuvent identifier les actions stratégiques à mener... Et prendre de meilleures décisions.

Un très bon exemple de ces interactions entre IA et analyse de données est **l'IA explicable (ou XAI)**. Ce concept fait référence à l'utilisation de techniques d'intelligence artificielle pour développer des modèles qui sont plus précis et efficaces pour faire des prédictions ou prendre des décisions. Et – cerise sur le gâteau – ces derniers sont également compréhensibles et

interprétables par les êtres humains (peu importe leur formation et connaissances en informatique).

Concrètement, l'IA explicable permet de **sortir le Big Data et l'IA de leur "boîte noire"**, en démocratisant les outils. Dans le domaine de l'analyse des données, le XAI peut offrir plusieurs avantages. Il aide notamment les Data Analysts à mieux comprendre comment et pourquoi un modèle d'IA particulier effectue certaines prédictions ou prend certaines décisions. Cela conduit à une plus grande confiance dans les résultats du modèle en particulier (et de l'IA en général).

Le XAI peut également **faciliter la validation, l'interprétation et la comparaison des modèles**. Cela permet aux analyses de données de mieux apprécier les performances d'un modèle, d'identifier certains biais ou limites... Et donc de prendre des décisions plus pertinentes non pas sur la base d'un modèle obscur, mais d'un modèle dont on comprend tous les tenants et aboutissants.

## **Comment l'apprentissage automatique peut-il contribuer à améliorer l'analyse des données ?**

On retrouve le même cercle vertueux entre machine learning et analyse de données. À bien des égards, **les analystes de données et les ingénieurs en apprentissage automatique se nourrissent** les uns les autres pour approfondir la compréhension de données.

Comme l'IA, le machine learning permet de **créer des modèles et des systèmes d'apprentissage automatique**, qui mettent à l'échelle de gros volumes de données, testent des hypothèses. Et pour finir, qui extraient des informations plus approfondies à partir de données jusque-là inexploitable.

En substance, l'apprentissage automatique va donc **compléter et optimiser le processus d'analyse de données** des manières suivantes :

### **Identifier des modèles à partir du Big Data**

Grâce à l'exploration des données et à la visualisation du Big Data, les analystes peuvent **identifier des modèles explicatifs ou prédictifs plus pertinents**. L'application d'algorithmes d'apprentissage automatique garantit en effet une compréhension plus complète et précise des modèles et tendances sous-jacents d'un gros volume de données.

### **Améliorer l'analyse prédictive**

Les modèles d'apprentissage automatique peuvent être entraînés de telle sorte qu'ils **produisent des prédictions plus fines et pertinentes**. Les entreprises peuvent ainsi limiter leur prise de risque, prévoir le comportement d'acteurs ou de marchés. Et surtout, prendre de meilleures décisions en conséquence.

### **Automatiser l'analyse de données**

Le machine learning permet aussi d'**automatiser les tâches à faible valeur humaine ajoutée** de la Data Analyse. On pense par exemple au nettoyage de données, à leur prétraitement ou encore à leur manipulation manuelle. Les Data Scientists peuvent en retour dédier plus de temps à l'interprétation et à la compréhension du Big Data.

## Renforcer la détection des anomalies

Les Data Analysts peuvent **confier la détection et la correction des erreurs aux algorithmes** d'apprentissage automatique. On le voit déjà dans le secteur financier ou celui de la cybersécurité, avec la détection des fraudes et virus.

## Mieux communiquer les résultats de l'analyse de données

Le machine learning améliore également la data visualisation. Intégré aux outils de visualisation, il offre **des représentations plus dynamiques et interactives** (en fonction de l'audience ou des besoins exprimés).

## Les défis à relever pour améliorer l'analyse des données grâce à l'IA et le machine learning

L'usage de l'IA et de l'apprentissage automatique dans la science des données a considérablement **amélioré et optimisé son impact et ses résultats**. Néanmoins, ce secteur d'activité présente encore de nombreux défis auxquels il faudra répondre, et notamment :

- **Le manque de données de formation.** Il est en effet particulièrement coûteux et fastidieux d'obtenir des données étiquetées pour former les modèles d'IA et de ML. L'apprentissage par transfert est l'une des méthodes les plus intéressantes pour résoudre ce problème. Il permet au modèle d'apprentissage automatique d'utiliser les informations utilisées pour réaliser les tâches ou créer les algorithmes précédents. Idem pour l'apprentissage auto-supervisé qui ouvre de nouvelles opportunités en matière d'usage de Big Data non étiquetée ;
- **L'évolutivité du modèle.** Les Data Scientists doivent en effet s'assurer que leurs algorithmes d'analyse soient suffisamment agiles et légers pour résister à l'épreuve du temps. Mais aussi, pour éviter qu'ils ne coûtent trop cher en maintenance.

# **En quoi consiste l'extraction de données ?**

## **Définition, processus et outils**

Si votre organisation peine à gérer toutes les informations qu'elle reçoit, sachez qu'elle n'est pas la seule.

Le volume d'informations qui arrive dans la plupart des organisations est en hausse. Qui plus est, les informations arrivent plus rapidement, par le biais d'une myriade de points d'entrée et dans une grande variété de formats et de types de fichiers. En règle générale, une organisation peut avoir besoin de saisir des informations à partir de documents scannés, de fichiers PDF, de formulaires PDF dynamiques, de feuilles de calcul Excel, de présentations PowerPoint, de documents Word, de formulaires électroniques et d'e-mails.

Pour couronner le tout, les informations reçues par les organisations sont plus complexes. Selon l'Association for Intelligent Information Management (AIIM), 63 % des informations reçues par une organisation moyenne ne sont pas structurées et, pour la plupart, ne sont pas gérées. Par conséquent, de nombreuses organisations ne sont pas en mesure d'identifier les métadonnées de base, telles que le propriétaire ou la version d'un document.

Sans visibilité claire sur les métadonnées, la conservation et l'élimination des documents deviennent alors des défis majeurs.

Le manque de visibilité sur les informations critiques complique également les processus essentiels dans l'ensemble de l'entreprise, notamment la correspondance avec les clients, les propositions de vente et les contrats, la gestion des dossiers, l'approvisionnement, les achats et les contrats, la recherche et le développement, la fabrication et l'entreposage, les ressources humaines, la finance et la comptabilité, ainsi que le service client du centre d'assistance.

La conjugaison de la pression incessante pour réduire les coûts, des exigences réglementaires et juridiques toujours croissantes et des attentes accrues en matière de services a fait de la saisie des données une priorité stratégique.

L'extraction de données remet de l'ordre dans le chaos de l'information. Lisez la suite pour découvrir comment.

## **En quoi consiste l'extraction de données ?**

L'extraction de données permet de saisir des données à partir de n'importe quelle source papier ou électronique en vue de leur traitement ou de leur stockage.

Dépassant largement les anciens systèmes de numérisation et de reconnaissance optique de caractères (OCR), les solutions modernes d'extraction de données sont basées sur l'IA, qui utilise l'apprentissage automatique supervisé pour classer et extraire intelligemment les métadonnées des documents structurés et non structurés, offrant ainsi plus de clarté et de rapidité.

Les informations saisies par ces solutions sont le plus souvent utilisées pour :

- La classification de documents
- La saisie de données textuelles
- Le transfert vers les processus
- L'indexation à des fins d'archivage
- L'indexation à des fins de récupération

Il s'agit d'un élément clé d'une stratégie de gestion de l'information.

## Comment fonctionne l'extraction de données ?

Comment ça marche :

1. **Entrée.** Les informations sont saisies à partir de scanners, de pièces jointes d'e-mails, d'appareils mobiles, de fax, de référentiels d'images et d'informations stockées.
2. **Classification.** Les solutions d'extraction de données permettent de classer et de comprendre les documents en utilisant le contexte et les mots uniques, les titres, les ancres clés, les valeurs, les modèles et les codes-barres.
3. **Extraire.** Cette technologie extrait des informations à l'aide d'un large éventail de technologies, notamment l'extraction non structurée de formes libres, l'extraction de codes-barres, l'extraction des textes manuscrits, l'extraction de détection de marque, l'extraction de bases de données « floues », l'extraction de PDF et de textes, et l'extraction de formulaires fixes.
4. **Transmission.** Les informations sont transmises aux systèmes en aval dans n'importe quel format, y compris XML, PDF, CSV, JSON, les référentiels de documents ou les interfaces de programmation d'applications (API).

Certains scanners de documents utilisent une technologie d'extraction de données intégrée pour traiter les documents mélangés, saisir les données pendant la numérisation, trier les exceptions et accélérer la configuration des documents. L'extraction des données pendant la numérisation des documents élimine également la nécessité d'indexer manuellement les types de documents et les métadonnées.

## Utilisations courantes et avantages

Bien que les logiciels d'extraction de données simplifient toutes les applications de gestion de l'information, cette technologie est particulièrement adaptée aux formulaires

semi-structurés et non structurés, tels que les factures, les documents hypothécaires et fiscaux, les dossiers de prêt, les demandes de souscription d'assurance, les formulaires d'assurance-maladie et les dossiers médicaux.

Prenez par exemple le back office d'une banque classique. L'équipe de gestion de l'information doit traiter les demandes de cartes de crédit, les formulaires d'ouverture de compte en ligne, les dossiers de prêts hypothécaires, les titres de propriété et autres documents de prêt, les chèques, les factures et les documents de remise de fonds. La plupart des banques s'appuient sur un ensemble hétéroclite de systèmes autonomes pour saisir les données de ces documents, ce qui se traduit par des processus inefficaces et coûteux.

Les salles de courrier peuvent utiliser ce logiciel pour mettre de l'ordre dans le volume toujours croissant de documents qu'elles reçoivent. La technologie uniformise, centralise et automatise la classification et la mise en lots des documents, la saisie des données figurant dans l'en-tête et les lignes, et l'exportation numérique des informations vers les systèmes et processus en aval, selon des règles d'entreprise préconfigurées.

Dans toutes ces applications, les solutions saisissent les informations au point de présentation.

De plus, les meilleures solutions d'extraction de données proposent une plateforme unique pour saisir les données des documents papier et numériques. Bien que le volume de documents papier soit en recul dans de nombreux segments, tels que le traitement des chèques et des explications des prestations, on ne sait pas quand le papier disparaîtra, ni même s'il disparaîtra un jour. En attendant, les organisations doivent trouver des moyens efficaces de fonctionner dans un environnement hybride de gestion de l'information, tout en respectant les exigences juridiques et réglementaires.

## Quels avantages en tirent les organisations

L'extraction de données change considérablement la façon dont les organisations saisissent les données.

- **Réduction des coûts.** L'extraction de données évite d'avoir à saisir manuellement les informations clés qui arrivent au sein d'une organisation, notamment les documents non structurés tels que les factures. Les métadonnées sont saisies avec une grande précision et l'apprentissage automatique améliore les performances au fil du temps. L'utilisation d'une plateforme unique pour saisir toutes les données permet aux organisations de consolider leurs systèmes.
- **Amélioration de la productivité du personnel.** Le temps que les employés perdent à saisir des données, à rechercher les documents papiers et les e-mails, et à corriger les erreurs est du temps qu'ils ne peuvent pas consacrer à des activités à plus forte valeur ajoutée, telles que l'analyse des données et la collaboration avec les principales parties prenantes. L'automatisation des

processus grâce à l'extraction de données permet au personnel de consacrer plus de temps aux activités les plus importantes.

- **Précision accrue.** Une simple faute de frappe peut engendrer de gros maux de tête en aval. Les solutions d'extraction de données utilisent l'IA pour saisir les informations avec un degré de précision élevé. Les données saisies peuvent être validées en les comparant aux informations contenues dans une application de gestion des ressources d'entreprise (ERP) ou dans un autre système d'archivage.
- **Temps de cycle réduits.** Les meilleures solutions de saisie de données utilisent l'automatisation des processus par la robotique (RPA) pour exporter en toute fluidité les données vers les systèmes en aval, sans qu'il soit nécessaire de les programmer. L'extraction des données au point de présentation garantit un flux plus rapide des documents.
- **Conformité simplifiée.** Les solutions d'extraction de données suivent toutes les actions effectuées sur un document. De plus, l'amélioration de la visibilité des métadonnées garantit le respect des délais de conservation des documents.
- **Évolutivité accrue.** Les solutions d'extraction de données permettent aux organisations de faire évoluer efficacement leurs opérations, sans avoir besoin d'engager du personnel supplémentaire lorsque leur volume augmente ou que leurs besoins changent.

Voici quelques-unes des raisons pour lesquelles de plus en plus d'organisations déploient des solutions d'extraction de données.

## Améliorez votre saisie de données

À l'ère du numérique, les informations parviennent aux organisations depuis une multitude de sources différentes, dans une variété de formats, structurés ou non. L'indexation et le balisage manuels sont des tâches lentes, sujettes aux erreurs et qui nécessitent trop de ressources. Les systèmes de reconnaissance optique de caractères (OCR) obsolètes n'offrent ni la rapidité ni les résultats dont les entreprises ont besoin. Les meilleures solutions d'extraction de données permettent aux entreprises de détecter les informations entrantes et d'en extraire des informations exploitables qui feront progresser l'entreprise.

# Le top 8 des meilleurs outils de data visualisation en 2024

## Qu'est-ce que la data visualization ?

La **data visualization** ou dataviz ou encore visualisation de données en français, c'est l'art de présenter ses données de la manière la plus claire et la plus engageante. Les méthodes de dataviz servent à prendre de meilleures décisions grâce à une meilleure compréhension des **données**, elles permettent également aux professionnels de gagner en conviction.

## Les principes de la visualisation de données

### 1) Comprendre son audience

Il s'agit certainement du principe le plus important, bien que les **visualisations de données** doivent être compréhensibles au plus grand nombre, les messages véhiculés diffèrent selon la personne adressée. En effet, vous devez **communiquer** différemment face à des inconnus, des collègues ou encore votre supérieur hiérarchique.

### 2) Se fixer des objectifs clairs

Afin que la présentation de vos données soit structurée, informative et convaincante, il est nécessaire de déterminer vos **objectifs**. Si, par exemple, vos données montrent une opportunité d'investissement en Île-de-France, votre exposé doit pouvoir convaincre de la pertinence de cet investissement.

### 3) Choisir les bons modèles de présentation

Une règle fondamentale et essentielle pour garantir la clarté et l'intégrité de vos **données**. Voici des recommandations générales pour le choix d'un modèle de présentation des données :

- Pour informer :
  - Un grand chiffre seul
  - Les **graphiques** donut
  - Les pictogrammes
- Pour comparer :
  - Les **bar charts**
  - Les bubble charts
  - Les mots en nuage
  - Les tree charts

- Les stack bar charts
- Pour suivre une évolution :
  - Les **timelines**
  - Les graphiques en courbe
  - Les cartes

Les pie charts et les effets 3D sont déconseillés pour la bonne compréhension des **données**. Pour certains projets, les graphiques ne sont pas utiles, n'ayez pas peur d'afficher uniquement de grands nombres ou seulement des **pictographes**.

#### 4) Simplicité

Une erreur de débutant consiste à créer des **visuels** complexes pour impressionner son manager. Si vous souhaitez que votre message soit compris, faites dans la simplicité. Ne surchargez pas vos visualisations d'informations inutiles.

#### 5) Intégrité

Si ce principe n'est pas respecté, vous courez peut-être à la catastrophe. Voici 4 conseils pour garantir la **véracité** de vos données :

- Afficher toutes les **données** pertinentes
- Utiliser les **échelles** correctement (commencer par 0 et garder la même unité de mesure tout le long du graphique)
- Une légende claire
- Éviter la confusion en choisissant les bons **visuels**

#### 6) Bien utiliser les couleurs

Il est important d'éviter de surcharger son **graphique** avec des couleurs trop nombreuses et variées. Il est souvent préférable d'opter pour un dégradé pour comparer les **données** entre elles.

#### 7) Fournir du contexte

Ne pas oublier d'ajouter du contexte pour que les **chiffres** soient compréhensibles. Cela inclut les titres, les légendes et parfois même des explications.

#### 8) Captiver son public

Un public captivé sera convaincu de la fiabilité de vos recommandations. La création d'un plan et l'utilisation de techniques de **storytelling** vous permettront de garder votre audience envoûtée.

## Les différents types d'outils de data visualisation

Passons au cœur du sujet, les outils de dataviz ! Avant de vous révéler notre **classement**, il est important de savoir que ces outils sont extrêmement variés. Plutôt que de créer un grand nombre d'articles sur chaque sous-famille de logiciels, nous avons décidé de regrouper les meilleures **technologies** en vous expliquant les différences de chacune d'entre elles.

### Les outils d'analytics

Ce sont les outils de data visualisation traditionnels. Lorsque l'on pense aux outils de **dataviz**, nous pensons rapidement à eux. Des outils comme **Tableau**, Looker Studio, Qlik Sense ou encore Power Bi.

Ces outils ont la particularité d'être des logiciels professionnels avec un nombre importants de fonctionnalités poussées allant même au-delà de la **visualisation de données**.

À l'aide d'une interface, vous pourrez visualiser et manipuler de grands **ensembles de données** sans connaissances en programmation. Ils sont facilement intégrables à diverses sources de **données** les rendant particulièrement intéressants pour les grandes entreprises.

### Les frameworks de programmation

Une technologie totalement opposée aux autres types. Les frameworks fonctionnent avec des **langages de programmation** du type Python, JavaScript ou encore R pour pouvoir créer des visualisations hautement performantes et personnalisables.

Programmer pour vos tâches de **data visualisation** comporte de nombreux avantages comme la possibilité de manipuler d'importants volumes de données rapidement, une grande flexibilité et un coût très faible. En revanche, cela nécessite une formation importante du langage de programmation et du **framework** associé.

### Les outils de design

Nous présentons peu ces logiciels comme des outils de **dataviz**. Pourtant, Canva ou Visme peuvent faire l'affaire si vous souhaitez créer de beaux visuels sans compétences et sans intégration de données.

Ces outils sont généralement abordables aussi bien techniquement que financièrement. Ils sont parfaits pour un usage privé ou en complément d'un **outil de visualisation de données** traditionnel.

# Comment choisir son outil ?

L'outil de data visualisation idéal diffère selon chaque organisation. Pour savoir lequel choisir, nous vous avons concocté une liste de **critères** à évaluer :

## Intégration des données

L'outil doit pouvoir intégrer le maximum de **sources** et de formats de données possibles, des fichiers excels aux systèmes de gestion de base de données. Sans une intégration fluide, vous risquez une perte significative d'informations.

## Fonctionnalités

Par « fonctionnalités », nous entendons un large choix de **modèles de visualisation** dont les diagrammes ou les graphiques, mais aussi la capacité de connecter des sources de données différentes ou encore de pouvoir les préparer. Les **fonctions** interactives sont aussi appréciées pour que vous puissiez explorer vos données en profondeur.

## Facilité d'utilisation

Un outil ergonomique et accessible vous fera gagner un temps considérable, il nécessitera, par ailleurs, peu de temps de formation et vous profiterez d'une **expérience d'utilisation** agréable.

## Collaboration

Pour le travail en équipe, cette feature est indispensable. Vous devriez pouvoir publier des rapports en ligne, partager des **tableaux de bord** avec des membres de votre équipe ou exporter des visualisations dans différents formats.

## Volume de données

L'outil doit être capable de gérer le **volume de données** avec lequel vous travaillez actuellement et à l'avenir. Vos performances ne doivent pas se dégrader pour ces raisons.

## Le prix

Le **coût** est aussi un facteur important, les politiques de prix varient grandement pour chacun de ces outils.

## Sécurité et gouvernance des données

Votre outil doit répondre aux normes de sécurité les plus strictes, surtout, si vous traitez des données sensibles ou propriétaires. Des fonctionnalités de **gouvernance** sont appréciables.

# Le top 8 des outils de datavisualization à choisir en 2024

## 1) Tableau

Le leader alliant performance et simplicité



Tableau est l'outil de **visualisation de données** leader sur le marché, il permet à tous types d'utilisateurs, la création de visualisations interactives et des tableaux de bord via une interface épurée et intuitive.

### Avantages

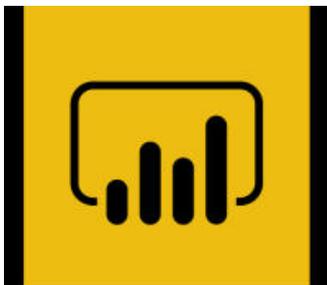
- Facile à prendre en main
- De nombreuses fonctionnalités pour la conception des **visualisations**
- Grande capacité d'intégration
- Communauté active et ressources abondantes pour apprendre

### Inconvénients

- Coût élevé, en particulier pour les petites entreprises et les utilisateurs individuels
- Peut nécessiter un ordinateur puissant pour gérer de grands ensembles de **données**

## 2) Power BI

Un outil très complet et intégrable à la suite Microsoft



Power BI est la suite Microsoft pour la **Business Intelligence**. Ce logiciel très complet permet de préparer, d'analyser des données et de concevoir des rapports puissants et diversifiés.

## Avantages

- Bonne intégration avec les produits Microsoft
- Prix compétitif, avec une version gratuite
- Capable de traiter de grandes quantités de **données**

## Inconvénients

- Manque de personnalisation pour les visualisations
- Une **interface** peu intuitive

## 3) Qlik Sense

Une alternative sérieuse à Tableau



Qlik Sense est un célèbre SaaS de data visualisation. Il utilise le Cloud pour pouvoir gérer de grands **flux de données**. Il est célèbre pour ses fonctionnalités de découverte guidée et d'exploration interactive.

## Avantages

- Forte capacité d'association des données, permettant une **exploration** intuitive
- Moteur en mémoire efficace pour un traitement rapide des données
- Interface conviviale et facile à utiliser

## Inconvénients

- Manque de **personnalisation** pour les visualisations
- Coût élevé, en particulier pour les petites entreprises et les utilisateurs individuels

## 4) Matplotlib et Seaborn

Pour les développeurs souhaitant une personnalisation maximale



Matplotlib et Seaborn sont des bibliothèques de **visualisation de données** Python. Elles permettent le développement rapide de graphiques hautement personnalisables.

## Avantages

- Grande personnalisation et grande diversité pour les visualisations
- Intégration simple avec l'écosystème Python
- **Gratuit**
- Open-source

## Inconvénients

- Complexe surtout pour les non-développeurs
- Le temps de création du fait de la nécessité de produire du **code**

## 5) Apache Superset

Le meilleur outil de data visualisation open-source



Apache Superset est un outil puissant, léger, gratuit et open-source. Via une interface épurée, vous pourrez l'utiliser pour produire de magnifiques **dashboards**. L'outil s'autohéberge, ce qui est un atout pour se conformer au RGPD.

## Avantages

- **Open-source**
- Gratuit
- **Interface utilisateur** plutôt intuitive et riche en fonctionnalités

## Inconvénients

- Très peu d'assistance en cas de problème
- Nécessité de posséder des compétences techniques pour l'installation et la configuration de l'outil

## 6) Canva

Le meilleur outil pour un usage simple et rapide



Canva est le leader des services de **design** en ligne. Son interface simple et épurée a conquis des millions d'utilisateurs à travers le monde. Il s'agit d'un outil très complet qui peut servir pour de simples **visualisations de données**.

## Avantages

- Très facile à utiliser, avec une **interface** en drag and drop
- Un large éventail de modèles et de designs prêts à l'emploi
- Convient aux utilisateurs sans compétences techniques
- Des **prix** compétitifs et une offre gratuite

## Inconvénients

- N'est pas adapté pour les visualisations complexes ou interactives.
- N'est pas adapté pour gérer de grands **volumes de données**

## 7) D3js

Pour les développeurs souhaitant concevoir des visualisations sur le web



D3js est une bibliothèque JavaScript pour produire des **visualisations** de données interactives sur un navigateur web.

## Avantages

- Grande personnalisation et grande diversité pour les visualisations
- Large communauté et de nombreuses ressources disponibles
- **Gratuit**
- Open-source

## Inconvénients

- Complexe, surtout pour les non-développeurs
- Le temps de création du fait de la nécessité de produire du code

## 8) Looker Studio

L'outil préféré du marketing pour l'analyse web



Looker Studio est l'outil de **dataviz** de Google, il permet de créer des présentations et des rapports à travers plusieurs sources de données, dont les services de la firme de Mountain View.

## Avantages

- Gratuit
- **Simple d'utilisation**
- Bonne intégration avec les produits Microsoft
- Possibilité de partager facilement les rapports et de collaborer

## Inconvénients

- Manque de personnalisation pour les **visualisations**
- Dépendance à l'écosystème Google
- Ralentissement lorsque de nombreuses sources de données sont utilisées en version gratuite

## Tableau Comparatif des meilleurs outils d'analytics

Critères	Tableau	Power BI	Qlik Sense	Apache Superset	Looker Studio
<b>Fonctionnalités</b>	Très avancées et variées	Très avancées et variées	Avancées, axées sur l'exploration de données et les visualisations interactives	Avancées	Avancées avec l'intégration de Google Cloud
<b>Volume de données</b>	Traite efficacement de grands volumes de données	Traite efficacement de grands volumes de données	Traite efficacement de grands volumes de données	Peut gérer de grands volumes de données, mais cela dépendra de l'infrastructure sur laquelle Superset est déployé	Traite efficacement de grands volumes de données en version payante
<b>Intégration des données</b>	Excellente, très large éventail de sources de données	Excellente avec une forte intégration des outils Microsoft	Grande quantité de sources de données	Grande quantité de sources de données, mais peut nécessiter des ajustements	Forte intégration des outils Google

Critères	Tableau	Power BI	Qlik Sense	Apache Superset	Looker Studio
<b>Facilité d'utilisation</b>	Une interface très intuitive	Assez simple d'utilisation	Une interface intuitive, mais, demande une courbe d'apprentissage modérée	Des compétences techniques sont nécessaires pour l'installation et la configuration de l'outil. L'interface est conviviale.	Interface intuitive, mais il est courant de rencontrer des problèmes de chargement en utilisant la version gratuite
<b>Collaboration</b>	Excellente, avec Tableau Server/Online	Excellente, intégration avec Teams	Bonne, grâce au cloud et au partage intégré	Variable, dépend de l'installation et de la configuration de l'entreprise	Très bonne, intégration avec les services Google
<b>Coût</b>	Modéré à élevé selon l'entreprise	Modéré, avantageux pour les entreprises utilisant les autres produits Microsoft	Modéré à élevé, se basant sur le nombre d'utilisateurs	Gratuit, mais il existe des coûts pour maintenir l'application hébergée	Une version gratuite et une version payante pour un usage avancé
<b>Sécurité et gouvernance des données</b>	Sécurité très élevée avec des options avancées pour la gouvernance des données	Sécurité élevée grâce à Azure Security, les outils de gestion renforcent la gouvernance des données	De très nombreuses options de sécurité et de gouvernance des données	Dépend de la manière dont l'outil est configuré ou déployé	Sécurité et option de gouvernance des données élevées grâce à Google Cloud

## Conclusion : Quel outil choisir ?

De manière générale, **Tableau** est l'outil préféré des professionnels. Très puissant et intuitif, il est souvent considéré comme étant la meilleure option.

Cependant, Tableau est un logiciel onéreux, si vous disposez d'un budget limité, des solutions autohébergées comme **Apache Superset** ou des frameworks de

programmation (si vous disposez de compétences en développement) pourront faire l'affaire.

En outre, Qlik Sense et Power BI sont des concurrents sérieux à Tableau. Si vous avez besoin de fortes capacités en exploration des données, Qlik Sense est pour vous. Par ailleurs, nous vous conseillons **Power BI**, si vous disposez d'un abonnement aux services Microsoft.

À l'instar de Power BI, Looker Studio gagne en intérêt si vous utilisez les services de Google. Cet **outil** pouvant rencontrer des problèmes de performance peut être utilisé gratuitement.

Enfin, si vous souhaitez créer de belles **visualisations** rapidement pour un besoin temporaire, vous pouvez utiliser le célèbre Canva.

# « Qu'est-ce que l'Internet des objets ? »



The screenshot shows the OVHcloud website header with navigation links like 'Mon compte client', 'Contact commercial', 'Webmail', 'Support', 'Communautés', and 'Learn'. Below the header, the article title 'Qu'est-ce que l'Internet des objets ?' is displayed. The main content area has a blue background and contains the following text:

**Qu'est-ce que l'Internet des objets ?**

Centre d'apprentissage | Qu'est-ce que l'Internet des objets ?

L'Internet des objets est souvent confondu avec la notion d'objet connecté. Système en pleine croissance, l'IoT est présent dans la vie domestique comme dans l'industrie, le commerce ou l'agriculture. Améliorer et automatiser les processus, optimiser la compréhension des besoins, le service client et faciliter la prise de décision : l'Internet des objets représente un progrès dans la vie quotidienne et valorise les entreprises. Envie d'en savoir plus ? Cette page vous permettra d'appréhender simplement la notion d'Internet des objets et d'en percevoir les avantages comme les inconvénients à travers des exemples parlants.



## Comprendre l'Internet des objets

### Internet des objets : définition

L'Internet des objets (ou IoT pour Internet of Things) désigne à la fois le processus de connexion d'objets physiques à Internet et le réseau qui relie ces objets.

Par « objets », on entend aussi bien des appareils de la vie quotidienne (domotique, montre de fitness, etc.) que des appareillages médicaux, des machines agricoles, des chaînes d'approvisionnement, des robots industriels ou des feux de circulation routière.

En définitive, l'IoT relie tout élément capable de transférer des données sur un réseau. Et ce, sans nécessiter d'interactions entre humains ou entre un humain et un ordinateur. Cependant, l'interaction personne-machine est rendue possible, ne serait-ce que pour procéder au paramétrage, à la configuration ou simplement pour accéder aux informations.

### Quelle est la différence entre objets connectés et Internet des objets ?

L'expression « objet connecté » englobe tout appareil connecté sans lien avec Internet (par exemple via le Bluetooth). L'Internet des objets, quant à lui, consiste en l'interconnexion entre le Web et les objets. Cependant, l'adjectif « connecté » est souvent utilisé par commodité pour désigner un appareil faisant partie de l'Internet des objets, comme une voiture connectée ou une maison connectée.

De plus, des personnes possédant un implant de moniteur cardiaque ou des animaux dotés d'une puce d'identification font également partie du réseau.

## Comment fonctionne l'IIoT ?

Un écosystème IIoT est composé d'objets compatibles avec le Web ou utilisant des systèmes informatiques embarqués.

Un objet connecté à l'IIoT est capable de collecter des données sur son environnement via des capteurs. Il exploite ensuite ces dernières via des processeurs avant de les envoyer à un ou plusieurs destinataires à l'aide de son matériel de communication intégré.

Il partage les données collectées par le biais d'une passerelle IIoT. Il s'agit d'une solution permettant la communication entre appareils ou entre ceux-ci et le cloud. Les informations sont donc transférées directement vers le cloud à des fins d'analyse et d'exploitation ou à un autre appareil IIoT pour une analyse locale.

Voici un exemple parlant pour illustrer ces propos. Lorsque vous approchez de votre habitation dans votre voiture connectée, celle-ci envoie vos informations de géolocalisation au thermostat du système de chauffage de la maison. En analysant ces données, ce dernier est à même de régler en votre absence la température intérieure de n'importe quelle pièce selon le paramétrage établi au préalable. L'intervention humaine n'a donc lieu qu'au moment de la configuration, l'IIoT des objets faisant le reste.

Bien entendu, les données récoltées sont exploitables en temps réel comme sur le long terme. In fine, les analyses peuvent être effectuées par des humains comme par une intelligence artificielle (IIA) disposant d'un système d'apprentissage automatique.

Ainsi, le système IIoT d'une maison connectée agit en temps réel pour déterminer le moment idéal du réglage du chauffage. Il peut également s'appuyer sur des données collectées par la voiture sur une longue période. De plus, toutes les données IIoT récoltées jour après jour par tous les thermostats connectés représentent une immense source de renseignements pour l'entreprise fournissant l'énergie. Elle peut en effet les analyser à des fins d'amélioration de ses services.

## Exemples d'IIoT qui révolutionnent notre quotidien

### IIoT grand public

De plus en plus d'appareils ménagers et d'objets domestiques sont dotés de capteurs et de systèmes de connexion dans le but d'améliorer le confort ou la sécurité (grille-pain connecté, bouteille de vin disposant d'un écran tactile et compatible avec le WiFi, etc.). Comme il est techniquement possible d'ajouter des capteurs IIoT à pratiquement n'importe quel objet du quotidien, l'IIoT grand public couvre un très grand nombre d'utilisations. Par exemple :

- les maisons intelligentes équipées de thermostats et de chaudières connectées, de systèmes d'éclairage intelligents et d'appareils électroniques connectés contrôlables à distance via un smartphone ou un ordinateur (ex. prises intelligentes, capteurs de mouvements, mangeoires pour animaux, home cinéma, machines à laver, vidéosurveillance, serrures, etc.) ;
- les voitures connectées améliorent le confort de conduite ainsi que la sécurité : climatisation, contrôle de vitesse, surveillance de la batterie et de la pression des pneus, emplacement du véhicule, ouverture automatique de la porte de garage ou du portail ;
- les appareils portables de santé ou de sport : capteur implantable, pompe à insuline, glucomètre, cardiofréquence-mètre, podomètre, compteur de calories, traceur GPS, etc.

A travers une application IoT, tous ces appareils peuvent fonctionner ensemble de manière harmonieuse et faciliter la vie quotidienne de leurs utilisateurs.

### **IoT d'entreprise et IoT industriel (IIoT)**

L'Internet des objets se démocratise dans tous les secteurs d'activité : production, transport, vente au détail, santé, agriculture, infrastructure, domotique, services publics, etc.

Dans le secteur de la santé par exemple, l'IoT a de nombreuses applications :

- surveillance fine des patients via une analyse en continu des données générées par des implants ou des capteurs ;
- gestion des stocks de produits et instruments et maintenance des appareils à l'hôpital ;
- suivi des signes vitaux des pompiers en intervention ou des ouvriers sur des sites à haut risque ;
- en cas d'urgence, calculs d'itinéraires optimisant le temps d'intervention des premiers secours, etc.

Dans le domaine de la sécurité, l'Internet des objets apporte de nombreuses solutions aux problématiques du contrôle d'accès et de l'authentification :

- capteurs connectés contrôlant l'entrée et la sortie des employés via un smartphone ;
- capteurs sur les machines-outils renforçant la sécurité au travail ;
- systèmes améliorant la sécurité des biens et des personnes dans les commerces ;
- détection des incendies, etc.

Dans le domaine de la fourniture de biens, les chaînes d'approvisionnement sont suivies et optimisées grâce aux capteurs et analyses IoT. Grâce au contrôle précis de la disponibilité d'une marchandise directement chez le consommateur et dans les entrepôts du fournisseur, le tout en s'appuyant sur les commandes en cours, un système IoT peut automatiquement déterminer l'opération de livraison la plus pratique pour les deux parties. En parallèle, le besoin de main-d'œuvre est optimisé, ce qui débouche sur une minimisation des coûts pour l'entreprise.

De manière générale, l'IoT permet d'améliorer la production, de réduire les temps d'immobilisation non planifiés des outils et des matières premières et d'augmenter la sécurité de très nombreux outils ou systèmes. Tous les secteurs industriels peuvent en bénéficier, avec des applications aussi diverses que la récupération des fibres dans l'industrie de transformation du bois ou le contrôle des foreuses sur une plateforme pétrolière.

La « ville intelligente », bardée de capteurs et d'applications IoT, constitue un parfait écosystème de l'Internet des objets : contrôle des émissions pour réduire la pollution de l'air, surveillance et contrôle du trafic des véhicules, économies d'énergie, etc. Elle comporte ainsi de multiples appareils et systèmes dits intelligents, allant du simple lampadaire à l'optimisation du trafic urbain, en passant par la vidéosurveillance ou la cartographie des émissions sonores.

Enfin, dans le secteur agricole, des systèmes utilisant l'IoT participent à la surveillance des champs et des serres de culture : luminosité, température, humidité de l'air et du sol, composition des sols, prévisions météorologiques, surveillance de la localisation et de la santé du bétail, etc. L'agriculture et l'élevage « intelligents » bénéficient également des ressources de l'IoT dans l'automatisation des systèmes d'irrigation des sols et d'alimentation du bétail.

En résumé, les possibilités d'application de l'Internet des objets sont quasiment infinies.

### **Avantages et inconvénients de l'IoT pour les entreprises**

#### **Quels sont les bénéfices de l'Internet des objets ?**

L'IoT offre de nombreuses possibilités, telles que :

- collecter des données sur n'importe quelle machine et y accéder depuis n'importe où, n'importe quand ;
- connecter de nombreux appareils électroniques en réseau et obtenir une communication optimale entre eux ;
- automatiser les tâches et processus ;
- améliorer la qualité des services et l'expérience client ;
- réduire la nécessité d'une intervention humaine et donc les coûts de main-d'œuvre ;
- gagner en temps et réduire les coûts sur les transferts de paquets de données ;
- améliorer la productivité et réduire les coûts de production ;
- surveiller et adapter l'ensemble des processus commerciaux ;
- optimiser la prise de décision, stratégique ou commerciale.

L'Internet des objets permet d'obtenir une vue en temps réel du fonctionnement d'un système en collectant, agrégeant et analysant des informations sur toutes ses composantes. Et ce, depuis la chaîne d'approvisionnement jusqu'aux opérations logistiques, en passant par la performance du parc d'équipements et des infrastructures.

#### **Quels sont les inconvénients de l'Internet des objets ?**

##### **Sécurité informatique et Internet des objets**

L'Internet des objets pose aujourd'hui des problèmes de sécurité des données et de confidentialité. En effet, l'IoT connecte des milliards d'appareils à Internet. Ceci nécessite par conséquent la sécurisation d'autant de points de données, chacun représentant un point d'attaque potentiel.

Un cybercriminel peut, en exploitant la vulnérabilité d'un seul point, profiter de l'étroite connexion entre les appareils IoT pour accéder à toutes les données du réseau. Il aurait ainsi la capacité de les voler, voire de les corrompre afin de les rendre inutilisables. La croissance du nombre d'appareils connectés et donc du volume de données échangées et partagées rend de fait le risque d'intrusion et de piratage de plus en plus important.

Ce risque est d'autant plus présent que la mise à jour des appareils IoT comblant notamment les failles de sécurité incombe souvent aux utilisateurs. Ces derniers ne les effectuant pas aussi régulièrement que nécessaire, le système devient par conséquent de plus en plus vulnérable.

La confidentialité des informations personnelles est elle aussi préoccupante :

- de nombreux objets connectés nécessitent la saisie par les utilisateurs de leurs informations de connexion, ou autres données personnelles, soumises aux risques de fuite ;
- les fabricants et distributeurs d'appareils IoT grand public peuvent utiliser ces derniers pour obtenir et vendre les informations personnelles des utilisateurs.

### **Autres inconvénients de l'Internet des objets**

Les failles de sécurité ne représentent pas le seul inconvénient de l'IoT :

- les entreprises en charge de la gestion et du maintien d'un parc très important d'appareils IoT doivent aussi faire face à une difficulté croissante en matière de collecte et de gestion des données issues de l'ensemble de l'écosystème IoT ;
- il n'existe pour l'heure pas de norme internationale de compatibilité dans ce domaine.

### **Quelle est l'histoire de l'Internet des objets ?**

Le premier objet connecté à Internet fut un distributeur de boissons fraîches à l'université Carnegie Mellon (États-Unis) en 1982. La notion d'appareil connecté existe, quant à elle, depuis les années 1970. À l'époque, l'expression « Internet des objets » n'est pas encore inventée : on parle alors d'Internet intégré ou d'informatique omniprésente.

Ce n'est qu'en 1999 que Kevin Ashton, informaticien chez Procter & Gamble, utilise pour la première fois l'expression « Internet des objets ». Cette appellation avait pour but de susciter l'intérêt des décideurs sur la technologie de radio-identification (RFID) ainsi que les autres capteurs qu'il souhaitait mettre en place sur les produits de leur chaîne d'approvisionnement.

Cette même année, le professeur à l'Institut de technologie du Massachusetts (MIT) Neil Gershenfeld publiait *When Things Start to Think*. Dans ce livre, il y décrit clairement la direction prise par l'interconnexion des machines, sans toutefois utiliser l'expression d'Ashton.

L'évolution de l'IoT s'est ensuite fortement construite sur la communication dite « machine to machine » (M2M). Elle fait référence aux appareils capables de se connecter les uns aux autres via un réseau, et ce, sans interaction humaine.

Aujourd'hui, l'IoT est un réseau de capteurs composé de milliards d'objets intelligents, connectant des personnes, des systèmes informatiques et des applications dans le but de partager et de collecter des données. L'IoT est rendu possible par la forte convergence technologique actuelle (Internet, technologies sans fil, systèmes microélectromécaniques, microservices, etc.) qui se situe entre technologie opérationnelle (OT) et technologie de l'information (IT).

# **Qu'est-ce que l'analyse des données ? Méthodes, techniques et outils**

## **Qu'est-ce que l'analyse des données ? Définition et exemple**

L'application systématique de techniques statistiques et logiques pour décrire la portée des données, moduler la structure des données, condenser la représentation des données, illustrer les données par des images, des tableaux et des graphiques, et évaluer les biais statistiques, les données de probabilité, afin de tirer des conclusions significatives, est connue sous le nom d'analyse des données. Ces procédures analytiques nous permettent d'induire l'inférence sous-jacente à partir des données en éliminant le chaos inutile créé par le reste des données. La génération de données est un processus continu, ce qui fait de l'analyse des données un processus continu et itératif dans lequel la collecte et l'analyse des données sont effectuées simultanément. Garantir l'intégrité des données est l'une des composantes essentielles de l'analyse des données.

Il existe plusieurs exemples d'utilisation de l'analyse des données : transport, détection des risques et des fraudes, interaction avec les clients, planification urbaine, soins de santé, recherche sur Internet, publicité numérique, etc.

Si l'on prend l'exemple des soins de santé, nous avons récemment observé qu'avec l'apparition de la pandémie de coronavirus, les hôpitaux doivent faire face à la pression du traitement d'un maximum de patients. L'analyse des données permet de surveiller l'utilisation des machines et des données dans de tels scénarios afin de réaliser des gains d'efficacité.

Avant d'aller plus loin, il convient de respecter les conditions préalables suivantes pour une bonne analyse des données :

- S'assurer de la disponibilité des compétences analytiques nécessaires.
- Veiller à l'application correcte des méthodes de collecte et d'analyse des données.
- Déterminer la signification statistique
- Vérifier si l'analyse est inappropriée
- Assurer la présence d'une inférence légitime et impartiale
- Garantir la fiabilité et la validité des données, des sources de données, des méthodes d'analyse des données et des déductions qui en découlent.
- Considérer la portée de l'analyse

## **Méthodes d'analyse des données**

Il existe deux méthodes principales d'analyse des données :

### **1. Analyse qualitative**

Cette approche répond principalement à des questions telles que « pourquoi », « quoi » ou « comment ». Chacune de ces questions est abordée au moyen de techniques quantitatives telles que des questionnaires, des échelles d'attitudes, des résultats standardisés, etc. Ce type d'analyse est généralement réalisé sous la forme de textes et de récits, qui peuvent également inclure des représentations audio et vidéo.

## 2. Analyse quantitative

En général, cette analyse se mesure en termes de chiffres. Les données sont présentées ici en termes d'échelles de mesure et sont étendues pour une manipulation statistique plus poussée.

D'autres techniques sont utilisées :

## 3. Analyse de texte

L'analyse de texte est une technique permettant d'analyser un texte pour en extraire des faits lisibles par une machine. Il vise à créer des données structurées à partir de contenus libres et non structurés. Le processus consiste à découper des piles de fichiers hétérogènes et non structurés en morceaux de données faciles à lire, à gérer et à interpréter. Elle est également connue sous le nom d'exploration de texte, d'analyse de texte et d'extraction d'informations.

L'ambiguïté des langues humaines est le plus grand défi de l'analyse de texte. Par exemple, les humains savent que « Red Sox Tames Bull » fait référence à un match de baseball, mais si ce texte est transmis à un ordinateur sans connaissances préalables, il générera plusieurs interprétations linguistiquement valides, et parfois les personnes qui ne s'intéressent pas au baseball pourraient avoir des difficultés à le comprendre également.

## 4. Analyse statistique

Les statistiques impliquent la collecte, l'interprétation et la validation des données. L'analyse statistique est la technique qui consiste à effectuer diverses opérations statistiques pour quantifier les données et appliquer l'analyse statistique. Les données quantitatives impliquent des données descriptives telles que des enquêtes et des données d'observation. Elle est également appelée analyse descriptive. Il comprend divers outils permettant d'effectuer des analyses statistiques de données, tels que SAS (Statistical Analysis System), SPSS (Statistical Package for the Social Sciences), Stat soft, etc.

## 5. Analyse diagnostique

L'analyse diagnostique est une étape supplémentaire de l'analyse statistique visant à fournir une analyse plus approfondie pour répondre aux questions. Elle est également connue sous le nom d'analyse des causes profondes, car elle comprend des processus tels que la découverte, l'extraction et le forage de données.

L'analyse diagnostique est une étape supplémentaire de l'analyse statistique qui permet d'approfondir l'analyse pour répondre aux questions. Elle est également connue sous le nom d'analyse des causes profondes, car elle comprend des processus tels que la découverte et l'extraction de données, ainsi que l'exploration et l'analyse approfondie.

Les fonctions de l'analyse diagnostique se répartissent en trois catégories :

1. **Identifier les anomalies** : après avoir effectué une analyse statistique, les analystes doivent identifier les domaines qui nécessitent une étude plus approfondie, car ces données soulèvent des questions auxquelles il est impossible de répondre en examinant les données.
2. **Approfondir l'analyse (découverte)** : l'identification des sources de données aide les analystes à expliquer les anomalies. Cette étape exige souvent des analystes qu'ils recherchent des modèles en dehors des ensembles de données existants et qu'ils extraient des données de sources externes, ce qui permet d'identifier des corrélations et de déterminer si l'une d'entre elles est de nature causale.
3. **Détermination des relations de cause à effet** : les relations cachées sont découvertes en examinant les événements qui ont pu conduire aux anomalies identifiées. La théorie des probabilités, l'analyse de régression, le filtrage et l'analyse des séries chronologiques de données peuvent être utiles pour découvrir des histoires cachées dans les données.

## 6. Analyse prédictive

L'analyse prédictive utilise des données historiques et les introduit dans le modèle d'apprentissage automatique pour trouver des modèles et des tendances critiques. Le modèle est appliqué aux données actuelles pour prédire ce qui va se passer ensuite. De nombreuses organisations la préfèrent pour ses divers avantages, tels que le volume et le type de données, les ordinateurs plus rapides et moins chers, les logiciels faciles à utiliser, le resserrement des conditions économiques et le besoin de différenciation concurrentielle.

Les utilisations courantes de l'analyse prédictive sont énumérées ci-dessous :

- **Détection des fraudes** : de multiples méthodes d'analyse améliorent la détection des schémas et préviennent les comportements criminels.
- **Optimisation des campagnes de marketing** : les modèles prédictifs aident les entreprises à attirer, conserver et développer leurs clients les plus rentables. Ils permettent également de déterminer les réactions ou les achats des clients, favorisant ainsi les opportunités de vente croisée.
- **Améliorer les opérations** : L'utilisation de modèles prédictifs concerne également la prévision des stocks et la gestion des ressources. Par exemple, les compagnies aériennes utilisent des modèles prédictifs pour fixer le prix des billets.
- **Réduction des risques** : le score de crédit utilisé pour évaluer la probabilité qu'un acheteur ne règle pas ses achats est généré par un modèle prédictif qui intègre toutes les données relatives à la solvabilité d'une personne. D'autres utilisations liées au risque comprennent les réclamations d'assurance et les recouvrements.

## 7. Analyse prescriptive

L'analyse prescriptive suggère diverses lignes d'action et décrit les implications possibles qui pourraient être atteintes après l'analyse prédictive. L'analyse prescriptive, qui génère des décisions ou des recommandations automatisées, nécessite une direction algorithmique spécifique, unique et claire de la part de ceux qui utilisent les techniques d'analyse.

## Processus d'analyse des données

Une fois que vous avez entrepris de rassembler les données à analyser, vous êtes submergé par la quantité d'informations que vous trouvez pour prendre une décision claire et concise. Avec autant de données à gérer, vous devez identifier les données pertinentes pour votre analyse afin d'en tirer une conclusion précise et de prendre des décisions éclairées. Les étapes simples suivantes vous aideront à identifier et à trier vos données en vue de leur analyse.

### 1. Précisez vos besoins en matière de données – définissez votre champ d'application :

- Définissez des questions courtes et directes, dont les réponses vous permettront de prendre une décision.
- Définir les paramètres de mesure.
- Définissez les paramètres que vous avez en tête et ceux que vous êtes prêt à négocier.
- Définissez votre unité de mesure. Par exemple : heure, monnaie, salaire et autres.

### 2. la collecte des données

- Collectez vos données en fonction de vos paramètres de mesure.
- Collectez des données à partir de bases de données, de sites web et de nombreuses autres sources. Ces données peuvent ne pas être structurées ou uniformes, ce qui nous amène à l'étape suivante.

### 3. Traitement des données

- Organisez vos données et n'oubliez pas d'ajouter des notes supplémentaires, le cas échéant.
- Recoupez vos données avec des sources fiables.
- Convertissez les données selon l'échelle de mesure que vous avez définie précédemment.
- Exclure les données non pertinentes.

### 4. Analyse des données

- Une fois que vous avez collecté vos données, effectuez le tri, le traçage et l'identification des corrélations.
- Au fur et à mesure que vous manipulez et organisez vos données, vous devrez peut-être revenir sur vos pas depuis le début, où vous devrez peut-être modifier votre question, redéfinir les paramètres et réorganiser vos données.
- Utilisez les différents outils disponibles pour l'analyse des données.

### 5. Déduire et interpréter les résultats

- Vérifiez si le résultat répond à vos questions initiales.
- Vérifiez si vous avez pris en compte tous les paramètres pour prendre votre décision.
- Vérifiez s'il existe des facteurs qui empêchent la mise en œuvre de la décision.
- Choisissez des techniques de visualisation des données pour mieux communiquer le message. Ces techniques de visualisation peuvent être des diagrammes, des graphiques, des codes de couleur, etc.

Une fois que vous avez une inférence, n'oubliez jamais qu'il ne s'agit que d'une hypothèse. Les scénarios de la vie réelle peuvent toujours interférer avec vos résultats. Dans le processus d'analyse des données, il existe certains termes connexes qui sont identifiés à différentes phases du processus.

#### 1. l'exploration de données

Ce processus implique des méthodes permettant de trouver des modèles dans l'échantillon de données.

#### 2. la modélisation des données

Il s'agit de la manière dont une organisation organise et gère ses données.

## Techniques d'analyse des données

Il existe différentes techniques d'analyse des données en fonction de la question posée, du type de données et de la quantité de données collectées. Chacun d'entre eux se concentre sur les stratégies d'acquisition de nouvelles données, d'extraction d'informations et d'approfondissement des informations pour transformer les faits et les chiffres en paramètres de prise de décision. En conséquence, les différentes techniques d'analyse des données peuvent être classées comme suit :

### 1. Techniques basées sur les mathématiques et les statistiques

- **Analyse descriptive** : L'analyse descriptive prend en compte les données historiques, les indicateurs clés de performance, et décrit la performance par rapport à une référence choisie. Elle tient compte des tendances passées et de la manière dont elles pourraient influencer les performances futures.
- **Analyse de dispersion** : Dispersion de la zone sur laquelle un ensemble de données est réparti. Cette technique permet aux analystes de données de déterminer la variabilité des facteurs étudiés.

- **Analyse de régression** : Cette technique fonctionne en modélisant la relation entre une variable dépendante et une ou plusieurs variables indépendantes. Un modèle de régression peut être linéaire, multiple, logistique, ridge, non linéaire, des données de vie et plus encore.
- **Analyse factorielle** : Cette technique permet de déterminer s'il existe une relation entre un ensemble de variables. Dans ce processus, il révèle d'autres facteurs ou variables qui décrivent les schémas de la relation entre les variables initiales. L'analyse factorielle fait un bond en avant dans les procédures de regroupement et de classification utiles.
- **Analyse discriminante** : Il s'agit d'une technique de classification dans l'exploration de données. Il identifie différents points dans différents groupes sur la base des mesures des variables. En termes simples, il s'agit de déterminer ce qui différencie deux groupes l'un de l'autre, ce qui permet d'identifier de nouveaux points.
- **Analyse des séries chronologiques** : dans ce type d'analyse, les mesures sont étalées dans le temps, ce qui nous donne une collection de données organisées connue sous le nom de série chronologique.

## 2. Techniques basées sur l'intelligence artificielle et l'apprentissage automatique

- **Réseaux neuronaux artificiels** : un réseau neuronal est un paradigme de programmation d'inspiration biologique qui présente une métaphore du cerveau pour traiter l'information. Un réseau neuronal artificiel est un système qui modifie sa structure en fonction des informations qui circulent dans le réseau. Les ANN peuvent accepter des données bruyantes et sont très précises. Ils peuvent être considérés comme très fiables dans les applications de classification et de prévisions commerciales.
- **Arbres de décision** : comme son nom l'indique, il s'agit d'un modèle d'arbre qui représente des modèles de classification ou de régression. Il divise un ensemble de données en sous-ensembles plus petits qui se développent simultanément en un arbre de décision connexe.
- **Programmation évolutive** : cette technique combine les différents types d'analyse de données à l'aide d'algorithmes évolutifs. Il s'agit d'une technique indépendante du domaine qui permet d'explorer un vaste espace de recherche et de gérer très efficacement l'interaction des attributs.
- **Logique floue** : il s'agit d'une technique d'analyse de données basée sur les probabilités qui permet de gérer les incertitudes dans les techniques d'exploration de données.

## 3. Visualisation et techniques graphiques

- **Diagramme à colonnes, diagramme à barres** : ces deux types de diagrammes sont utilisés pour présenter les différences numériques entre les catégories. Le graphique en colonnes prend la hauteur des colonnes pour refléter les différences. Les axes sont intervertis dans le cas du diagramme à barres.
- **Graphique linéaire** : ce graphique est utilisé pour représenter l'évolution des données sur un intervalle de temps continu.
- **Graphique de surface** : Ce concept est basé sur le graphique linéaire. En outre, il remplit de couleur la zone située entre la polyligne et l'axe, représentant ainsi une meilleure information sur les tendances.
- **Graphique circulaire** : Il est utilisé pour représenter la proportion des différents classements. Il ne convient que pour une seule série de données. Cependant, il peut être composé de plusieurs couches pour représenter la proportion de données dans différentes catégories.
- **Graphique en entonnoir** : Ce graphique représente la proportion de chaque étape et reflète la taille de chaque module. Cela permet de comparer les classements.
- **Graphique nuage de mots** : il s'agit d'une représentation visuelle de données textuelles. Elle nécessite une grande quantité de données, et le degré de discrimination doit être élevé pour que les utilisateurs perçoivent les plus saillants. Ce n'est pas une technique d'analyse très précise.
- **Diagramme de Gantt** : montre le temps réel et la progression de l'activité par rapport aux exigences.
- **Graphique radar** : utilisé pour comparer plusieurs graphiques quantifiés. Il représente les variables dans les données qui ont des valeurs plus élevées et celles qui ont des valeurs plus faibles. Un graphique en radar est utilisé pour comparer le classement et les séries, ainsi que la représentation proportionnelle.

- **Nuage de points** : montre la distribution des variables sous forme de points dans un système de coordonnées rectangulaires. La distribution des points de données peut révéler la corrélation entre les variables.
- **Graphique à bulles** : il s'agit d'une variante du diagramme de dispersion. Ici, en plus des coordonnées x et y, la surface de la bulle représente la 3ème valeur.
- **Graphique** : Il s'agit d'une sorte de graphique matérialisé. Ici, l'échelle représente la métrique, et le pointeur représente la dimension. Il s'agit d'une technique adaptée à la représentation des comparaisons d'intervalles.
- **Diagramme de cadre** : Il s'agit d'une représentation visuelle d'une hiérarchie sous la forme d'une structure arborescente inversée.
- **Diagramme en arbre rectangulaire** : Cette technique est utilisée pour représenter des relations hiérarchiques mais au même niveau. Il permet une utilisation efficace de l'espace et représente la proportion que représente chaque zone rectangulaire.
- **Carte régionale** : utilise la couleur pour représenter la distribution des valeurs sur une partition de la carte.
- **Carte à points** : Représente la répartition géographique des données sous forme de points sur un fond géographique. Lorsque les points sont de la même taille, cela n'a aucune signification pour les données individuelles, mais si les points sont en forme de bulles, cela représente en plus la taille des données dans chaque région.
- **Carte des flux** : représente la relation entre une zone d'entrée et une zone de sortie. Elle représente une ligne reliant les centres de gravité géométriques des éléments spatiaux. L'utilisation de lignes de flux dynamiques permet de réduire l'encombrement visuel.
- **Carte thermique** : représente le poids de chaque point dans une zone géographique. La couleur représente ici la densité.

## Outils d'analyse des données

Il existe plusieurs outils d'analyse des données sur le marché, chacun ayant ses propres fonctions. La sélection des outils doit toujours être basée sur le type d'analyse effectuée et le type de données traitées. Vous trouverez ci-dessous une liste de quelques outils convaincants pour l'analyse des données.

### 1. Excel

Il dispose d'un grand nombre de fonctionnalités intéressantes et, avec l'installation de plugins supplémentaires, il peut traiter une quantité massive de données. Ainsi, si vous disposez de données qui sont loin de la marge significative de données, alors Excel peut être un outil très polyvalent pour l'analyse des données.

### 2. Tableau

Il entre dans la catégorie des outils de BI, conçus dans le seul but d'analyser les données. L'essence même de Tableau est le tableau croisé dynamique et le graphique croisé dynamique, qui permettent de représenter les données de la manière la plus conviviale possible. Il dispose également d'une fonction de nettoyage des données ainsi que de brillantes fonctions analytiques.

### 3. BI Power

Initialement conçu comme un module externe pour Excel, il s'en est ensuite détourné pour devenir l'un des outils d'analyse de données les plus importants. Il existe en trois versions : Free, Pro et Premium. Ses langages PowerPivot et DAX permettent de mettre en œuvre des analyses avancées sophistiquées similaires aux formules d'Excel.

### 4. Rapport détaillé

Fine Report est doté d'une simple opération de glisser-déposer, qui permet de concevoir différents styles de rapports et de construire un système d'analyse décisionnel des données. Il peut se connecter

directement à toutes sortes de bases de données, et son format est similaire à celui d'Excel. En outre, il fournit également une variété de modèles de tableaux de bord et plusieurs bibliothèques de plug-ins visuels développés par ses soins.

## **5. R et Python**

Il s'agit de langages de programmation très puissants et flexibles. R est le meilleur pour l'analyse statistique, comme la distribution normale, les algorithmes de classification en grappes et l'analyse de régression. Il effectue également des analyses prédictives individuelles telles que le comportement des clients, leurs dépenses, les articles préférés des clients en fonction de leur historique de navigation, etc. Elle fait également appel à des concepts d'apprentissage automatique et d'intelligence artificielle.

## **6. SAS**

Il s'agit d'un langage de programmation pour l'analyse et la manipulation des données, qui permet d'accéder facilement aux données de n'importe quelle source. SAS a lancé un ensemble complet de produits de profilage des clients pour l'analyse du web, des médias sociaux et du marketing. Vous pouvez prévoir leurs comportements, gérer et optimiser les communications.

## **Conclusion**

Il s'agit d'un guide complet pour les débutants sur le thème « Qu'est-ce que l'analyse des données ? L'analyse des données est la clé de toute entreprise, qu'il s'agisse de lancer un nouveau projet, de prendre des décisions de marketing, de poursuivre un plan d'action particulier ou de procéder à un arrêt complet. Les déductions et les probabilités statistiques calculées à partir de l'analyse des données permettent d'éclairer les décisions les plus critiques en éliminant le biais humain. Les différents outils d'analyse ont des fonctions qui se chevauchent et des limites différentes, mais ils sont aussi des outils complémentaires. Avant de choisir un outil d'analyse des données, il est essentiel de prendre en compte l'étendue du travail, les contraintes d'infrastructure, la faisabilité économique et le rapport final à préparer.

# Descriptif du SI de Commaggio

**Le SI métier est composé de plus de 15 applicatifs. Certaines applications métiers sont hébergées dans un environnement Cloud**

- L'infrastructure cloud utilise AWS, Azure et Google Cloud pour permettre une scalabilité immédiate

## 1. Matériel (Hardware)

- Serveurs :
  - Des serveurs rackables (Dell PowerEdge R740) sont utilisés pour héberger les applications, les bases de données et les serveurs web.
  - Un serveur de fichiers (NAS Synology) pour le stockage et le partage de fichiers sécurisé.
- Stockage :
  - SAN (Storage Area Network) EMC Unity pour la gestion des données à grande échelle.
  - NAS (Network Attached Storage) : Synology DS220+ pour le stockage de fichiers.
- Équipements réseau :
  - Routeurs et commutateurs : routeurs Cisco ISR 4000 et des commutateurs Cisco Catalyst 9000 pour gérer le trafic réseau.
  - Pare-feu : Un dispositif de sécurité Fortinet FortiGate pour protéger le réseau interne.
- Postes de travail :
  - 900 postes, ordinateurs portables et de bureau (Dell XPS et Dell OptiPlex)
- Périphériques :
  - Imprimantes réseau (HP) et scanners pour la gestion documentaire.

## 2. Logiciel (Software)

- Systèmes d'exploitation :
  - Windows Server pour les serveurs.
  - Linux (CentOS) pour certains serveurs d'applications.
- Environnement de travail individuel :
  - Microsoft 365 pour les outils de collaboration et de productivité.
  - Logiciels de gestion de projet Trello
- Bases de données :

- SQL Server, MySQL et Oracle pour la gestion des données.
- Outils de virtualisation :
  - VMware pour créer des machines virtuelles, optimisant l'utilisation des ressources

### **3. Réseau (Network)**

- Infrastructure de câblage :
  - Câblage Ethernet CAT6 pour les connexions réseau.
  - Points d'accès Wi-Fi (Cisco pour la connectivité sans fil.
- VPN (Virtual Private Network) :
  - Pour permettre aux agents de se connecter de manière sécurisée depuis l'extérieur.

### **4. Sécurité (Security)**

- Antivirus et antimalware :
  - Solutions Norton pour protéger les appareils contre les menaces.
- Systèmes de sauvegarde :
  - Logiciels de sauvegarde Veeam pour assurer la récupération des données.
- Contrôle d'accès :
  - Gestion des identités et des accès (IAM) via la solution Microsoft Azure AD.

### **5. Centre de données (Data Center)**

- Un environnement climatisé avec des racks pour les serveurs, des systèmes d'alimentation sans interruption (UPS) pour garantir la disponibilité.
- Surveillance de l'infrastructure à l'aide de systèmes de gestion d'infrastructures de centre de données (DCIM).