LA LETTRE D'INFORMATIONS
mc² iNews
GROUPE

HORS SÉRIE #9

# DATA & ANALYTICS

Octobre 2018

**Alexis Trentesaux**
**Directeur de l'Expertise Data & Analytics**

L'année 2018 marque les 30 ans du premier raccordement à Internet en France. Qui aurait pu imaginer à cette époque la quantité d'informations qui allait parcourir chaque jour le monde, la vitesse à laquelle nous pouvons actuellement accéder à ces données et l'émergence du participatif au travers du Web 2.0.

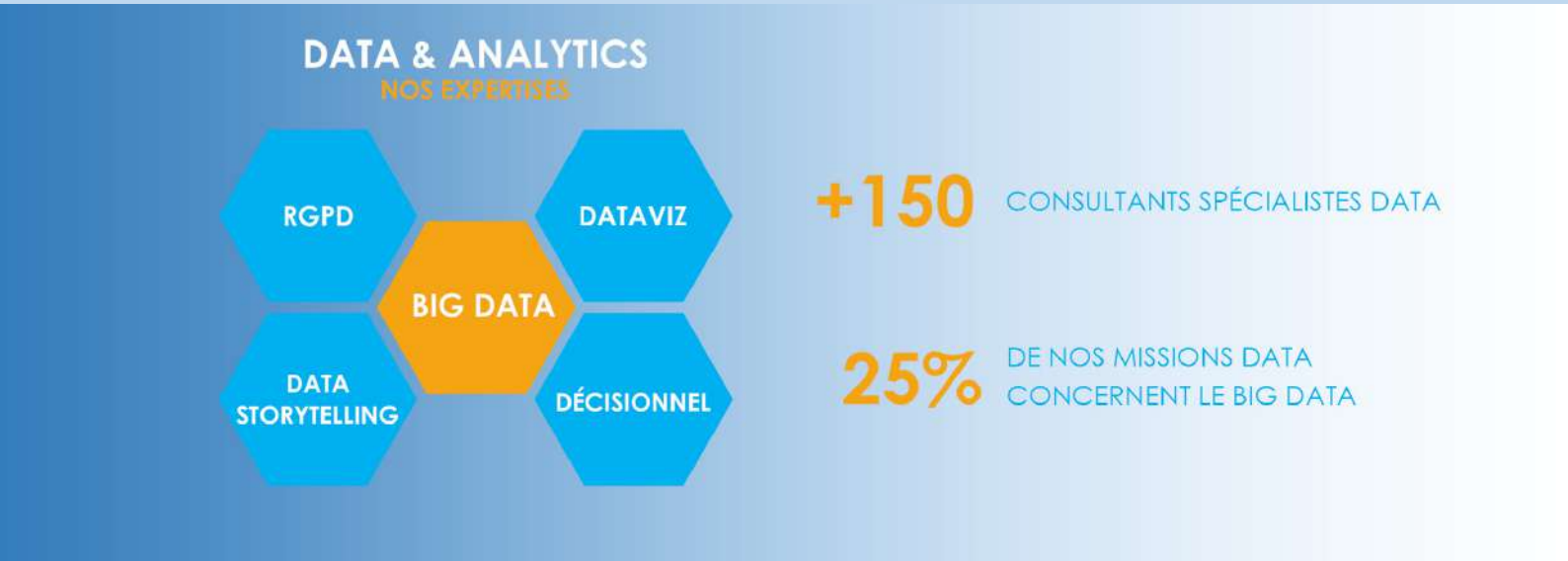
L'évolution des usages, tant pour les particuliers que pour les entreprises, met aujourd'hui la donnée au centre des préoccupations des sociétés avec la volonté de vouloir restituer et analyser toujours plus d'informations auprès d'un maximum de personnes, tout en assurant une qualité optimale des process dans l'entreprise.

Désormais le partage d'information doit se faire de manière industrialisée, automatisée, mais surtout sécurisée et réfléchie car 2018 est aussi un tournant dans le monde de l'informatique avec l'entrée en vigueur du RGPD.

Enfin, il est important de se projeter vers de nouveaux usages avec des approches toujours plus disruptives permettant de repenser des métiers.

La blockchain, l'analyse augmentée et l'Open Data sont des sujets moteurs qui donneront toujours plus de dynamisme à nos entreprises et d'intérêt à nos métiers.
Réjouissons-nous donc de ces innovations qui ont marqué ces trente dernières années et gardons cet esprit de pionner qui nous permettra de tracer une route vers de nouveaux horizons…

A travers cette Newsletter Hors-Série Data & Analytics, nous allons traiter divers sujets tels que : le Process Mining, le Big Data Analytics As a Service ou encore les Smart Data. Je vous souhaite une bonne lecture.



# PROCESS MINING, DE L'EXCELLENCE OPÉRATIONNELLE AU RGPD

Article mc² Groupe publié dans SOLUTIONS

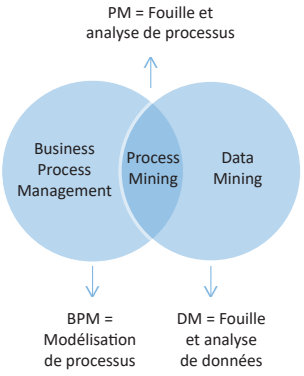
Dans un environnement concurrentiel et en pleine mutation, les entreprises de tous les secteurs doivent aujourd'hui se réinventer pour survivre.

Grâce aux capacités croissantes des Systèmes d'Information, la course à la performance opérationnelle s'en trouve facilitée : il est aujourd'hui devenu possible pour une entreprise d'introduire des assistants intelligents et automatisés au sein même de ses procédés. Il s'agit d'outils de Process Mining, qui exploitent les empreintes numériques laissées par chaque instance de processus opérationnel, dans le but de permettre à l'organisation d'optimiser ses performances. Mais au-delà de la course à l'excellence opérationnelle, le Process Mining peut-il également être utilisé au service de la mise en conformité avec le RGPD ?

**Process Mining = BPM + Data Mining**

Le Process Mining est l'une des dernières évolutions du Business Process Management (ou BPM). Cette technique analytique permet de dresser un état des lieux complet, objectif des procédés opérationnels, tels qu'ils se sont réellement déroulés, dans le but de les comprendre, les maîtriser et les améliorer.

Pour arriver à ses fins, le Process Mining, allie les disciplines d'exploration de la donnée (Data Mining) et de modélisation et analyse de procédés (BPM), comme montré dans le schéma ci-dessous.




*Définition du Process Mining*

En effet, le Data Mining est la discipline qui permet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, en utilisant des méthodes automatiques ou semi-automatiques. Le Business Process Management est quant à lui une démarche permettant de visualiser et de fluidifier les processus opérationnels de l'entreprise dans une optique d'amélioration de sa performance.

Si nous retenons l'extraction d'un savoir à partir de la donnée, spécifique au Data Mining, et l'observation et l'amélioration des processus opérationnels de l'entreprise, spécifiques au BPM, nous obtenons une démarche permettant d'analyser les données et métadonnées du déroulement réel des processus opérationnels de l'entreprise dans une optique d'amélioration de sa performance. En d'autres termes le Process Mining !

A noter que le Process Mining, contrairement au Data Mining et au BPM, puise ses racines en grande partie dans les métadonnées et plus précisément dans les « event logs », des traces d'exécution laissées par les processus informatiques dans les SI et généralement inexploitées par les entreprises.

**Entreprises, visez l'excellence opérationnelle !**

La fouille de procédés se déroule généralement en trois étapes. La première est la découverte automatisée des procédés ou Business Process Discovery. Il s'agit de la visualisation et la découverte des processus tels qu'ils se sont réellement déroulés, selon un niveau de complexité généralement paramétrable. Cette étape permet une meilleure connaissance globale du ou des processus étudiés.

La deuxième étape est la vérification de la conformité ou Conformance Checking. Il s'agit de la comparaison des processus réalisés avec des modèles théoriques ou inversement, afin de constater et documenter les écarts entre le prévu et le réalisé. Elle permet d'avoir un aperçu du degré de conformité des processus aux règles d'entreprise.

La troisième et dernière étape est l'amélioration des procédés (Business Process Improvement) qui consiste à identifier, grâce aux graphiques et visuels obtenus, des sources d'erreurs (ex : goulot d'étranglement), sources d'inefficiences (ex : sous-exploitation des ressources), risques potentiels, mais surtout des sources d'amélioration et d'optimisation des processus.

Ces étapes, bien qu'elles se suivent chronologiquement, peuvent néanmoins être considérées individuellement comme des objectifs à part entière. Si elle est utilisée à bon escient, la fouille de procédés peut même avoir un quatrième objectif : permettre de simuler des processus et de prédire leurs comportements futurs. Grâce à la connaissance des faits réels et à l'identification des bons et mauvais élèves parmi les processus de l'entreprise, il est possible de comprendre les origines de leur réussite ou échec, et d'identifier les leviers à actionner pour simuler une performance opérationnelle hors pair.

Le Process Mining, au-delà de ses principaux objectifs décrits ci-dessus, présente plusieurs avantages pour les organisations.

Tout d'abord, cette technique est non-intrusive et transparente pour les procédés et les équipes. Comme elle se base sur des données opérationnelles réelles (souvent des métadonnées), elle ne nécessite aucune création de données ou saisie d'informations complémentaires. De plus, contrairement aux audits de processus classiques qui nécessitent du temps et de l'énergie et mobilisent une multitude d'acteurs, le Process Mining est un ensemble d'algorithmes automatisés. Les workshops, entrevues et autres ateliers de travail avec les différents salariés concernés cèdent alors la place aux algorithmes et automatismes.

Un autre avantage majeur des technologies de PM est son indéniable objectivité, grâce à la traçabilité (informatique) de la donnée. Les faits, chiffres et event logs ne mentent pas et les perceptions subjectives n'y ont pas leur place.

Enfin, un dernier avantage mais pas le moindre, est l'aspect particulièrement visuel des outils de Process Mining. Grâce aux représentations graphiques et à la cartographie de l'organisation et des processus informatiques fournis par ces outils, la compréhension du fonctionnement réel des procédés s'en trouve simplifiée et à la portée de tous.

**Mise en conformité au RGPD : une nouvelle application du Process Mining ?**

A l'approche de la date du 25 mai 2018, l'acronyme RGPD (GDPR en anglais) est revenu régulièrement dans l'actualité des entreprises et a occupé les éditeurs de logiciels qui proposent de plus en plus de solutions et modules informatiques permettant de faciliter et d'accompagner la mise en conformité.

" Un autre avantage majeur des technologies de PM est son indéniable objectivité, grâce à la traçabilité informatique de la donnée. "

Le RGPD concerne le traitement des données à caractère personnel. Or, qui dit traitement de données… dit aussi processus. Comment alors ne pas faire un lien avec le Process Mining ? Et si le Process Mining permettait de répondre, du moins en partie, aux contraintes de conformité imposées par l'Union Européenne ?

Les technologies de Process Mining permettent de maîtriser le cycle de la vie de la donnée, en reconstituant de bout en bout son parcours informatique, et en fournissant un mapping visuel du traitement qu'elle subit, depuis son origine (arrivée dans le SI) jusqu'à sa destination finale (qu'il s'agisse d'un stockage, d'un archivage ou d'une purge). Cette cartographie est une étape primordiale de la mise en conformité au RGPD, grâce à laquelle l'organisation pourra vérifier si les critères légaux de base du traitement de la donnée sont respectés. La CNIL conseille d'ailleurs aux entreprises de commencer par « recenser de façon précise les traitements de données personnelles que vous mettez en œuvre » [1]. Pourquoi ne pas utiliser les technologies du Process Mining pour ce faire ?

***mc²i Groupe propose un accompagnement global permettant de répondre à ces différents enjeux pour une mise en conformité à la fois rapide et maîtrisée en matière de risques.***

**Mise en conformité**

- Cartographie des données
- Construction du registre des traitements
- Etude de l'impact sur la vie privée (PIA)
- Constitution de la documentation et des procédures
- Mise à jour juridique des contrats
- Pilotage du projet de mise en conformité

[1] *https://www.cnil.fr/fr/cartographier-vos-traitements-de-donnees-personnelles*

# LES SMART DATA DANS L'ANALYSE PRÉDICTIVE AU SEIN DE LA SANTÉ

Article mc²i Groupe publié dans VILLE intelligente

Les Smart Data et l'Intelligence Artificielle permettent de prendre des décisions commerciales ou de comparer nos performances sportives et sont de plus en plus utilisées dans le domaine de la santé pour anticiper les risques et apporter des soins personnalisés et adaptés aux patients.

**Vers une médecine sur-mesure ?**

D'où proviennent ces données médicales ? Tout d'abord, il y a les données qui existaient avant l'informatisation des usages, comme par exemple le dossier médical des patients, détenues par les institutions et les professionnels de la santé.

Avec l'arrivée du Big Data, de nouvelles données sont apparues via les réseaux sociaux, les objets connectés ou notre cyber environnement. Grâce à ces données, la personnalisation des soins est devenue un enjeu. Le groupe Vitalia utilise les informations de ses patients pour : « répondre à leurs besoins spécifiques, notamment pour la préparation de l'hospitalisation et la convalescence ». La fédération Unicancer a quant à elle mis en place une solution d'analyse sémantique des dossiers de ses patients, afin d'aider les médecins dans le choix du traitement le plus adapté.

Quant aux données des études médicales, elles peuvent être biaisées car subjectives. En effet, pour la plupart des études, ce sont directement les patients qui rapportent leurs comportements. L'utilisation d'objets connectés et d'applications pourrait résoudre ce problème en aidant les utilisateurs à suivre leur progression vers un mode de vie plus sain. Le suivi et la surveillance de l'activité physique, d'une alimentation équilibrée mais également des maladies chroniques sont ainsi facilités.

**Le futur de la santé 2.0 : la médecine prédictive**

Des programmes comme IBM Watson existent et sont désormais utilisés dans le cadre de diagnostics. Les algorithmes s'avèrent désormais plus efficaces que l'analyse humaine pour détecter des cancers. Dans le cas de patients souffrant d'obésité, l'accompagnement grâce à des objets connectés est très efficient pour éviter la reprise de poids.

Le Japon est un pays précurseur en la matière et annonce les prémices d'une véritable révolution dans l'utilisation de la robotique pour les soins. Les robots sont utilisés pour surveiller les patients isolés ou pour prodiguer des soins à distance. De même, des robots de « compagnie » aident à apaiser les patients atteints d'Alzheimer.

Le secteur de la santé est en passe de subir une transformation profonde grâce aux Smart Data et à l'Intelligence Artificielle. Si les robots sont encore loin de remplacer les professionnels de la santé, ces technologies bouleversent déjà les usages des patients et du corps médical.

**La santé à l'épreuve du RGPD**

Depuis le 10 avril dernier, le Système National des Données de Santé est entré en service. Il s'agit d'une base de données centralisant toutes les informations de santé des Français. La FMF, syndicat professionnel de médecins libéraux français, alerte sur « le risque élevé de perte de confidentialité des données personnelles des patients ».

Le volume de données de santé est une source précieuse de savoirs pour accompagner le malade, mais la protection de la vie privée des patients et le respect du secret médical impliquent d'accroître la protection et de maîtriser l'utilisation de ces données.

Les données de santé étant personnelles, confidentielles et sensibles, il est primordial d'identifier l'usage final de la donnée recueillie ainsi que sa durée de conservation. L'entrée en vigueur de la nouvelle règlementation RGPD en mai 2018 permet notamment de disposer d'un cadre juridique commun. Le principal défi reste d'assurer la confidentialité en anonymisant les données tout en garantissant la sécurité et la conformité de ces données.

" Si les robots sont encore loin de remplacer les professionnels de la santé, ces technologies bouleversent déjà les usages des patients et du corps médical. "

Les Smart Data sont en train de transformer considérablement l'avenir de la médecine. Elles permettent de dresser un portrait objectif d'un patient, de lui offrir des soins sur mesure et d'aider les médecins dans le choix du traitement à prescrire. Mais avec le Big Data et l'aide de l'IA, il devient possible de prédire certaines maladies et de les prévenir avant qu'il ne soit trop tard. Ces possibilités ne doivent néanmoins pas cacher le côté éminemment personnel et confidentiel de telles données. L'éthique des usages doit être pensée bien en amont.

# L'ÉMERGENCE DE L'ANALYSE AUGMENTÉE


Article mc²i Groupe publié dans Silicon.fr

Bien que souvent déjà matures avec la Business Intelligence (BI) intégrée à leurs processus métiers, plusieurs entreprises se sont dotées de technologies innovantes autour du Big Data telles que l'IA et le Machine Learning afin d'améliorer leurs processus d'affaires.

Selon IBM, 90% des données mondiales ont été créées durant ces deux dernières années. D'ici 2025, une multiplication par 8 du volume total de données à analyser est attendue. Un volume important qui défie toujours plus la capacité d'analyse de l'homme.

Dans cette course mondiale l'économie de la data, comment exploiter efficacement l'ensemble des données dont nous disposons.

**Intelligence Artificielle & Business Intelligence : une combinaison gagnante ?**

Lorsque l'on nous parle d'Intelligence Artificielle, on pense au fameux test de Turing de 1950 visant à créer une intelligence capable de reproduire la conversation humaine. Mais l'IA depuis a bien changé. Son fondement reste le même : créer des machines capables de simuler l'intelligence, mais ses cas d'usage se sont diversifiés (Chatbot, Machine Learning…), notamment dans la prise de décision.

La Business Intelligence est un ensemble d'outils et de méthodes visant à transformer des données brutes en connaissances afin d'aider la prise de décision. La BI que nous connaissons aujourd'hui traite principalement d'importants volumes de données structurées.

Dans l'analyse augmentée, terme introduit par Gartner, la BI complétée par l'IA permet d'exploiter l'ensemble des données non structurées à disposition, favorisant ainsi l'analyse complexe (notamment prédictive) pour les entreprises.

L'IA n'en reste cependant pas là. L'automatisation des traitements grâce au Machine Learning, permet également aux humains de se dégager plus de temps pour des tâches à plus forte valeur ajoutée. La majorité des organisations s'avèrent sensibles aux promesses de l'IA puisque 8 entreprises sur 10 en 2017 ont déjà commencé à investir dans un marché mondial estimé à 37 milliards de dollars d'ici 2025. Qlikview (éditeur de solutions BI) rajoute même que : « L'Intelligence Artificielle n'est pas strictement artificielle. L'approche idéale amplifie l'intuition humaine à l'aide d'une intelligence informatique pour créer un puissant effet multiplicateur. C'est l'intelligence augmentée. ».

**Comment procéder ?**

Pour toute analyse de données, l'essentiel repose dans la préparation et le traitement des données en amont. En BI, ces données sont collectées, nettoyées, transformées et stockées dans un entrepôt de données grâce à un ETL, puis analysées et présentées grâce à des outils de restitutions (Business Object, Tableau, Qlik).

" L'instauration d'une bonne gouvernance de la donnée est donc un élément central et le développement d'une culture de la donnée au sein de l'entreprise facilitera amplement son essor. "

Dans une combinaison de l'IA et de la BI, la préparation des données s'apparente fortement à celle de la BI plus classique. Cependant, l'apport complémentaire majeur de l'IA repose sur l'intégration des données non structurées grâce à l'analyse du langage naturel, et sur une automatisation de l'ensemble de ces tâches, souvent réalisées par l'Homme, à l'aide du Machine Learning.

En effet, l'IA pourrait en amont proposer de collecter des données externes, de les nettoyer et de les associer entre elles par hiérarchie, par type de données ou encore par regroupement. Il en va de même pour le partage des données pour lequel la manipulation et la présentation pour les décideurs en sont simplifiées puisque restituées à travers des outils qui leur sont déjà familiers. L'IA pourrait directement leur proposer une analyse afin de privilégier et favoriser des prises de décisions plus rapides et effectuées avec plus de confiance.

Néanmoins, le point déterminant d'une bonne décision résulte surtout de la qualité des données qui impactera nécessairement la qualité de l'analyse selon le principe du « Garbage In, Garbage Out ». La précision des prédictions est par conséquent intrinsèquement liée à la précision des données en amont. L'instauration d'une bonne gouvernance de la donnée est donc un élément central et le développement d'une culture de la donnée au sein de l'entreprise facilitera amplement son essor.

Enfin, cette combinaison des capacités cognitives de l'Homme, de la Business Intelligence et de l'Intelligence Artificielle semble être la formule idéale pour accroître considérablement le potentiel d'exploitation de nos données.

# DATA ANALYTICS : L'ENJEU STRATÉGIQUE DU 21ÈME SIÈCLE


A une époque où la quantité des données ne cesse de croître, la prise de décision n'a jamais autant été influencée par la manière dont nous interprétons ces données.

Data Analytics, une expression galvaudée à l'ère de l'information, mais dont la signification varie en fonction des contextes. Les méthodes d'analyse qui lui sont associées sont nombreuses, leurs objectifs différents et leurs applications infinies. Levons ensemble le voile sur cette brique cruciale dans la valorisation des « Big Data ». La Data Analytics, concrètement qu'est-ce que c'est ?

**Données, informations, connaissance**

Nous utilisons la Data Analytics au quotidien, lorsque nous lisons, écoutons, observons. Alors que les données nous submergent, et que notre capacité à les synthétiser est limitée, il nous est nécessaire de les stocker, de les traiter et de les analyser en masse. C'est pourquoi nous recourrons à l'informatique.

L'analyse de données (« Data Analytics ») correspond à l'interprétation que nous donnons à une réalité perçue partiellement (les « données ») afin d'en extraire du sens (« l'information »), puis de la cohérence (« connaissance »).

Elle regroupe un ensemble de méthodes qui permettent de synthétiser des informations inexploitables a priori par l'esprit humain, du fait de leur nombre ou de leur complexité.

La « donnée » est directement issue d'une mesure. Elle fait référence à la perception particulière d'un fragment de réalité. « L'information » quant à elle est une donnée ou un ensemble de données disposant d'un sens ou à laquelle nous donnons une interprétation. La donnée peut être exploitée dans le cadre d'une décision et ce n'est qu'à ce stade qu'elle acquiert réellement sa valeur.

**Quels cas d'usage pour les Data Analytics ?**

De nombreuses organisations entreprennent d'analyser quotidiennement des données. Les marketeurs exploitent les données dont ils disposent sur leurs clients pour mieux les comprendre et ainsi proposer des produits personnalisés. Certains sites internet ont par exemple recours au « Revenue Management », une technique qui consiste à estimer le prix qu'un acheteur serait prêt à payer à partir de l'historique internet de son navigateur.

En raison des enjeux commerciaux ou politiques de notre perception de la réalité, plusieurs critères sont cruciaux au cours d'une analyse.

- Le premier est la qualité des données exploitées, comme par exemple la précision d'une mesure.

- Le second est la représentativité : est-ce qu'une mesure régulière de la température dans une agglomération sur une année me suffit pour analyser l'évolution du climat terrestre ?
- Le troisième point est la méthodologie utilisée.
- Enfin, il est nécessaire de rester prudent quant à notre capacité à appréhender le monde au travers de nos analyses : c'est toute la différence entre corrélation et causalité.

Après tout, une analyse ne produit qu'une représentation ou une modélisation synthétique et partielle de la réalité. Son objectif principal est de rendre cette analyse la plus exhaustive possible. Mais comment procéder ?

**Analyse qualitative et quantitative**

L'analyse de données est découpée de manière générique en deux grandes familles : l'analyse qualitative et l'analyse quantitative.

La première exploite des données non structurées, mais lisibles et interprétables par l'humain (transcriptions d'entretiens clients, recherches documentaires…). Le recours à l'analyse qualitative est généralement l'une des premières étapes d'un protocole de recherche en sciences humaines ; les chercheurs l'utilisent notamment afin de construire des hypothèses qui seront ensuite testées via une approche quantitative.

Cet autre type d'approche utilisera des technologies variées afin de produire une synthèse d'un plus grand nombre de données. Il est fondé sur les mathématiques et est implémenté dans un système informatique, du fait du nombre d'opérations que son utilisation requiert. Aujourd'hui, ce type d'analyse est de plus en plus poussé. Il devient même possible d'expliquer les données « en continu », grâce notamment aux algorithmes d'apprentissage : chaque nouvelle observation est intégrée au modèle quantitatif, dans une logique d'amélioration continue.

Si l'on souhaitait à réaliser la segmentation de la clientèle des produits de beauté, nous pourrions :

- Procéder à une analyse qualitative en faisant passer un entretien de 30 minutes à chacun de nos clients ; une démarche coûteuse et surtout irréalisable avant que les informations obtenues ne soient obsolètes.
- Emettre des hypothèses sur les facteurs principaux du comportement d'un individu, collecter les informations sur ces facteurs principaux (via un questionnaire par exemple) auprès d'une large population, puis analyser ces données en masse à l'aide d'outils statistiques. C'est l'analyse quantitative.

**Statistiques descriptives vs statistiques inductives**

La statistique est la branche des mathématiques qui étudie les données et leurs relations entre elles. Elle se divise en deux catégories distinctes : les statistiques descriptives et les statistiques explicatives (ou inductives).

L'objet de la statistique descriptive est de synthétiser les données considérées : il s'agit de recourir à un ensemble de mesures qui permettront de se faire une idée de la population observée. De nombreux indicateurs relèvent de la statistique descriptive : moyenne, médiane, écart-type… Il s'agit ici d'extraire de « l'information » à partir des « données ».

Pour autant, afin de pouvoir prendre une décision, il n'est parfois pas suffisant de pouvoir décrire les données en elles-mêmes. Nous recourrons alors au raisonnement inductif afin de faire émerger une théorie à partir de l'observation de faits. Disposer des données météorologiques enregistrées depuis un siècle pourrait permettre d'établir une tendance et d'anticiper les variations climatiques. C'est l'objet des statistiques inductives.

La règle que l'on essaie d'induire à partir des données est représentée par un modèle. Dans le cadre d'une modélisation statistique, l'observateur va tenter de trouver le modèle qui explique le mieux ces données.

" Afin de pouvoir prendre une décision, il n'est parfois pas suffisant de pouvoir décrire les données en elles-mêmes. "

Les modèles statistiques sont des objets probabilistes, c'est-à-dire qu'ils laissent place à un certain degré d'incertitude. Tout l'art de la statistique inductive réside dans la capacité à trouver un modèle qui laisse le moins de place possible à l'incertitude, tout en étant une représentation fidèle de la réalité. L'objectif de l'estimation du modèle est de déterminer les caractéristiques de cette loi. Ce modèle simple pour lequel la marge d'erreur peut être grande, ne permet pas de prendre une décision. En effet, je ne peux pas décider des vêtements que je porterai demain uniquement en me basant sur la température qu'il fait aujourd'hui sans m'exposer à un coup de froid imprévu. Nous chercherons donc un modèle plus complexe, qui fera intervenir d'autres paramètres.

Bien entendu, afin d'augmenter la précision d'un modèle, autrement dit son degré d'adéquation avec la réalité, il est nécessaire d'augmenter le nombre d'observations qui servent à l'estimer. C'est pourquoi la prolifération des données, associée aux solutions techniques permettant de les traiter – le fameux Big Data - participent largement à renforcer notre capacité prédictive.

**Quelle importance pour les entreprises ?**

L'objectif général de l'analyse de données est de fournir une synthèse précise de la réalité ou des mécanismes qui la sous-tendent, appréhendable par l'esprit humain, à partir d'un nombre réduit d'observations. Cette brique est cruciale dans le processus de décision, comme l'a prouvé l'actualité récente : la victoire de Donald Trump aux élections présidentielles aurait été influencée par « Cambridge Analytica ». Cette entreprise aurait utilisé un modèle comportemental nommé « OCEAN », qui permet d'estimer le comportement d'un ensemble d'individus à partir des « likes » de leur compte Facebook.

Ce genre d'exemple nous prouve que l'application de la Datascience a énormément gagné à l'apparition de l'informatique et à la prolifération des données. Leur utilisation dispose plus que jamais d'une valeur ajoutée considérable dans la prise de décision stratégique.

# BIG DATA ANALYTICS AS A SERVICE : LE FUTUR DE L'ANALYSE DE DONNÉES DANS LE NUAGE

Article mc² Groupe publié dans IT for BUSINESS

Alors que le Big Data acquiert une place prépondérante dans notre éco-système, avec ses nombreux cas d'application (santé, économie, marketing, énergie…), force est de constater que les entreprises ne sont pas toutes séduites par l'ensemble des nouvelles technologies que son exploitation requiert.

Pour autant, beaucoup sont attirées par les potentielles retombées financières de la transformation en temps réel des données en information. C'est la raison pour laquelle le marché du Big Data Analytics as a Service (BDAaaS) se développe de plus en plus.

**Pourquoi une solution orientée Cloud ?**

Durant les trois dernières décennies, les nouvelles technologies de la révolution informatique se sont développées de manière exponentielle, obligeant les entreprises à faire face à un choix déterminant : investir régulièrement dans leur SI ou conserver leurs architectures existantes.

Or, ce choix est à la fois financièrement difficile et largement incertain : des investissements lourds ne signifient pas nécessairement un résultat proportionnel, surtout en sachant qu'au début des années 2010, plus de 50% des projets informatiques étaient considérés comme des échecs au regard du triptyque Qualité-Coût-Délai. D'un autre côté, ne pas franchir certains caps technologiques peut s'avérer dangereux, comme l'a prouvé l'ancien géant de la photographie Kodak, qui a raté le tournant du numérique.

Une troisième possibilité a émergé dans les années 2000 : les solutions 'as a Service', qui reposent sur l'externalisation partielle d'un SI auprès d'un éditeur chargé de la maintenance, de la sécurité et de l'architecture physique (et parfois logicielle).

Ces Softwares ou Platforms as a Service (SaaS ou PaaS) rencontrent actuellement un succès notable, et le développement de nombreuses briques « aaS » est fulgurant : on parle aujourd'hui de « Database as a Service », d'« Infrastructure as a Service », de « Backend as a Service » ou encore de « Games as a Service ».

**Quel en est l'intérêt à l'ère du Big Data ?**

Une architecture Big Data se compose de différentes briques, avec des problématiques spécifiques à chacune : diversité des sources pour la collecte, performance pour le stockage, puissance de calcul pour l'analyse, et adaptation des modes de restitution pour le reporting. Ceci constitue une première difficulté rencontrée par les entreprises lors de leur adoption du Big Data.

D'autres difficultés complexifient la mise en place de solutions Big Data, parmi lesquelles la recherche de profils qualifiés, le bon choix de la solution à implémenter et la décision d'investir les ressources nécessaires tout en acceptant un ROI incertain.

La solution « aaS » est sans doute l'une des meilleures réponses que le marché puisse apporter à ces différentes problématiques : externalisation des profils, solution testable, investissement limité et souplesse en cas de coupure budgétaire.

" Ces Softwares ou Platforms as a Service (SaaS ou PaaS) rencontrent actuellement un succès notable, et le développement de nombreuses briques « aaS » est fulgurant. "

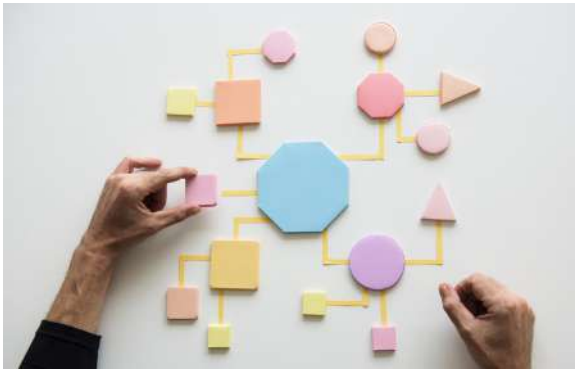
Les éditeurs doivent cependant affronter un challenge de taille, qui est de fournir une solution accessible et opérationnelle à des utilisateurs parfois peu matures. Ils doivent également faire face à des problématiques structurelles, telles que gérer la diversité et la qualité des données, assurer leur sécurité et leur confidentialité et fournir une gamme variée et maitrisée de services.

**Quels sont les acteurs du Big Data Analytics as a Service ?**

A ce jour, les services Big Data Analytics clé en main forment une population largement hétérogène. La Data Discovery Platform de Wipro, lancée en 2017, basée sur la plateforme Microsoft Azure, dispose par exemple d'un business model spécifique : le Pay-per-Insight. Mais d'autres acteurs plus récents se développent également, comme Emcien fondée en 2002 ou encore Arcadia Data fondé en 2012, qui proposent des solutions d'analyse et de restitution des données quasiment en temps réel.

Pour autant, les perspectives d'évolution de ce marché sont vastes. Outre le marché explosif estimé par IDC à plus de 203 milliards de dollars en 2020 contre 130 en 2017, les freins à l'adoption de leur propre plateforme Big Data persistent pour les entreprises. Les géants du Web tels que Google, Facebook, Microsoft et IBM ont tout intérêt à investir massivement sur le marché de l'Analytics ; déjà maîtres des technologies, infrastructures et compétences sous-jacentes, la monétisation de l'information à l'ère du Big Data est et sera une source majeure de croissance dans les années à venir.

# PROJET BIG DATA : POURQUOI LA MÉTHODE CRISP EST-ELLE INCONTOURNABLE ?



Article mc²i Groupe publié dans Solutions Magazine

Depuis la révolution numérique et l'explosion du volume de données, nombreuses sont les entreprises qui engagent des projets Big Data. Meilleur traitement des données, meilleurs bénéfices, nouvelles offres et nouveaux marchés… les opportunités sont alléchantes. Mais le tableau est sombre.

Parmi ces projets innovants, seuls 15% atteindraient la phase d'industrialisation post-POC (Proof of Concept), comme le montre une étude du Gartner datant de Juillet 2016 . Les raisons principales de cet échec ? Une indéniable inadéquation entre les attentes (parfois surréalistes) des métiers et la réalité du Big Data et une mauvaise maîtrise de ces nouvelles technologies. Heureusement, le modèle vertueux du CRISP apporte quelques solutions aux obstacles qui entravent la réussite de ces projets.

**CRISP : Une méthode en avance sur son temps**

Le CRoss-Industry Standard Process ou CRISP était initialement dédié à l'exploration de données ou Data Mining (CRISP-DM, DM pour « Data Mining »). Aujourd'hui généralisée à tous types de projets Big Data, cette intrépide méthodologie résistante aux vingt-et-une années passées depuis sa création est toujours de loin la méthodologie la plus populaire dans la mise en place de ces projets.

Le consortium CRISP-DM attribue en 2000 le succès de sa méthodologie à la conduite de projets pratiques d'exploration de données pendant plus de deux ans et demi : « CRISP-DM réussit parce qu'il est fondé sur l'expérience pratique et réelle de la façon dont les gens mènent des projets d'exploration de données ».

Loin donc de la théorie, la méthodologie CRISP est également totalement indépendante de l'industrie technologique puisqu'elle impose un schéma standard applicable à tout type de projet et/ou d'infrastructure.

Par ailleurs, cette méthodologie a la particularité d'adopter une démarche cyclique et itérative, permettant une meilleure appréhension des spécificités de chaque projet, comme le montre l'illustration ci-dessous.

**Démarche cyclique et itérative de la méthode CRISP**

CRISP ne propose pas un chemin linéaire unique entre le démarrage du projet et son déploiement. A contrario, la marche arrière est non seulement autorisée, mais recommandée : les ajustements en cours de route sont les bienvenus, puisqu'ils permettent de garder le modèle efficace.

**Des aspects métiers à ne pas négliger**

« Si j'avais une heure pour sauver le monde, je passerais 59 minutes à définir le problème et 1 minute à trouver la solution ».

Cette célèbre phrase d'Albert Einstein démontre ce que de nombreuses méthodologies de projet n'arrivent malheureusement pas à appliquer aujourd'hui : la résolution d'un problème devient un jeu d'enfant dès lors que les enjeux métier en sont clairement définis.

En effet, contrairement aux idées reçues, le Big Data n'est pas un ensemble abstrait de technologies complexes. Bien qu'il utilise les apports de la Data Science pour valoriser la donnée, il puise son origine dans un besoin métier préalablement identifié, s'inscrit dans un schéma business en accord avec la stratégie globale de l'entreprise.

" Cette méthodologie a la particularité d'adopter une démarche cyclique et itérative, permettant une meilleure appréhension des spécificités de chaque projet. "

C'est pour cette raison que CRISP constitue un excellent cadre pour la mise en place d'un projet Big Data : cette méthodologie accorde une attention particulière à la définition du besoin et des objectifs métiers, ainsi qu'à la collaboration continue entre le métier et l'équipe projet.

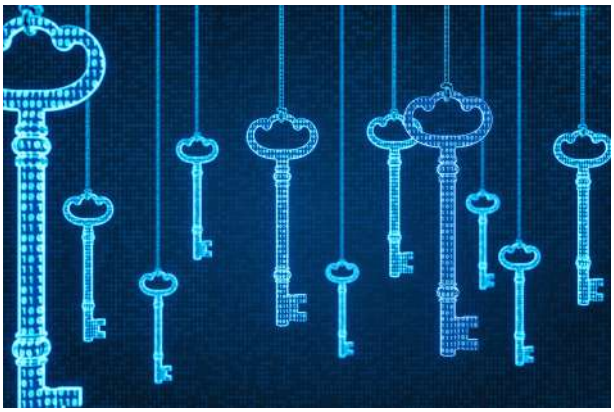
**Une évolution vers plus d'agilité**

Bien que prônant le modèle itératif et les nombreuses possibilités de retours en arrière, force est de constater que la méthode CRISP peut dans certains cas manquer d'agilité.

Dans un monde ivre d'incertitudes, où le besoin métier évolue rapidement, la mise en place d'un dialogue régulier entre les équipes projets et les équipes techniques devient indispensable. Ainsi, en complément du mode itératif caractéristique du CRISP, l'association d'un framework agile permettra une plus grande réactivité face aux changements. Scrum, Kanban, Lean, aucune méthodologie agile n'est incompatible avec CRISP. Chacune des six étapes de la méthodologie CRISP peut alors être assimilée à une user story.

Aucune méthodologie n'est cependant parfaite, et la clé de la réussite résidera toujours dans l'implication constante des métiers pour une amélioration continue du produit final. Après tout, comme disait le statisticien George E.P.BOX "All models are wrong, but some are useful» *(traduction : Tous les modèles sont faux mais certains sont utiles).*

# LA BLOCKCHAIN ET LE BIG DATA POUR UNE NOUVELLE RÉVOLUTION



Pour de nombreux spécialistes, la Blockchain est l'outil qui révolutionnera le numérique. Les gouvernements et les grandes entreprises, notamment du secteur Banque et Assurance, y voient déjà le moyen de modifier en profondeur leur organisation et de sécuriser leurs transactions de façon optimale. Si pour certains domaines, l'application de cette technologie est une évidence, elle l'est un peu moins pour d'autres ; c'est le cas du Big Data.

**Comment la Blockchain adresse-t-elle les problématiques du Big Data ?**
Les GAFA (Google, Apple, Facebook et Amazon) ont été les précurseurs dans la création et l'utilisation d'outils innovants permettant la gestion de fortes volumétries de données. Des projets open source ont suivis et ont permis le développement d'outils de traitement des données en temps réel.

Cependant, les pratiques liées aux traitements des données par les technologies du Big Data soulèvent de nombreuses interrogations, qui sont mises en lumière notamment lorsqu'on installe une application sur son smartphone : qui est propriétaire des données utilisateurs ? Comment les utilisateurs clients peuvent-ils savoir dans quel but et de quelle manière elles seront utilisées ? Qui peut y avoir accès ? Sont-elles anonymes ou nominatives ?

La technologie Blockchain est à même d'apporter une réponse à ces questions, et de replacer les utilisateurs clients au centre du système, en les laissant choisir les informations qu'ils acceptent de dévoiler. Prenons l'exemple du Bitcoin et de la circulation des informations sur sa Blockchain.

Le Bitcoin utilise les propriétés de la cryptographie asymétrique afin de permettre à n'importe quel utilisateur de crypter une transaction via une clé publique et d'en décrypter une autre qui lui est adressée via sa clé privée. Cette technique permet ainsi la protection et l'anonymat des données des utilisateurs.
Par conséquent, nous pouvons imaginer le développement d'algorithmes de type « Machine Learning » permettant une réelle évolution du modèle Big Data tout en garantissant la protection des données personnelles des utilisateurs. Par exemple, les données finales résultant de ces études pourront être anonymisées et ainsi cibler des « profils type » plutôt que des personnes.

Ceci ouvrirait la porte à un nouveau modèle d'infrastructure : un stockage décentralisé garantissant l'immuabilité des données via la Blockchain et un traitement distribué des données via les outils de l'écosystème Big Data.

**Comment lier Blockchain et architecture Big Data ?**
Les points majeurs qui ralentissent actuellement le développement des Blockchains au niveau des entreprises sont la performance et la scalabilité de la technologie : nombre de transactions par seconde limité, latence d'une opération d'écriture, copie intégrale de la Blockchain et mise à jour en temps réel sur chaque nœud.

" La technologie Blockchain est à même d'apporter une réponse à ces questions, et de replacer les utilisateurs clients au centre du système, en les laissant choisir les informations qu'ils acceptent de dévoiler. "

Ces performances sont bien en dessous de celles des bases de données NoSQL modernes qui, elles, peuvent excéder plusieurs centaines de milliers de transactions par seconde et offrir des stockages de plusieurs pétabytes avec une latence de l'ordre d'une fraction de seconde.

C'est sur ce constat que le projet BigchainDB a mis en place une base de données distribuée et scalable, capable d'adapter la puissance nécessaire en fonction du besoin. Pour y parvenir, elle utilise la Base de données MongoDB (première base de données en NoSQL) à laquelle s'ajoute une couche technologique supplémentaire, celle de la Blockchain. Cela permet ainsi de bénéficier de ses avantages tels que le contrôle décentralisé, le transfert d'actifs digitaux et la garantie de disposer de données sécurisées.

Mais alors, comment la Blockchain peut-elle modifier en profondeur l'usage d'une base de données de type Big Data ?
A l'inverse d'une base de données classique de grande envergure, il n'est plus besoin de se demander si l'information partagée est à jour, si elle a été retraitée et par qui. Que ce soit pour une société ayant des bureaux à travers le monde, ou une entreprise partageant des données avec d'autres entreprises, la Blockchain permettra à tous les utilisateurs de disposer d'un même niveau de contrôle sur leurs données, via un contrôle et un accès décentralisés, de tracer chaque transaction/information, chaque traitement effectué et enfin d'identifier l'acteur concerné.

Au travers de ce nouveau cas d'utilisation, la Blockchain illustre une fois de plus son potentiel d'application et les gains importants qu'elle permettrait de réaliser.

# OPEN DATA : COMMENT OUVRIR VOS BASES DE DONNÉES ?

Il y a 2 ans, le 7 octobre 2016, la « loi pour une République numérique » a été adoptée. Parmi les mesures phares : la mise en ligne par défaut des données publiques, l'Open Data. L'objectif est de libérer l'accès aux données publiques, pour permettre plus de transparence et stimuler l'innovation. Cette libéralisation a un coût non négligeable. Mais les enjeux sont également importants avec un marché européen de l'ordre de 30 milliard d'euros.

**Diffusion et gouvernance**

Quel intérêt pour les entreprises privées ? Bien plus réticentes à donner accès à leurs données, elles ont pourtant tout intérêt à le faire : innovation, productivité... Mais ces projets nécessitent une réflexion en amont.

Avant de lancer tout projet, il convient d'identifier le vecteur de diffusion qui sera utilisé donc de comprendre le type de données dont on dispose. Deux options peuvent être envisagées :

- Le premier mode de diffusion se fait sous la forme de fichier brut, en format standardisé téléchargeable directement sur un portail : il s'agit du mode de diffusion le plus facile à mettre en place mais également du plus difficile à maintenir.
- L'alternative est l'utilisation d'API, un protocole de communication entre le serveur de données et le client. Plus complexe à mettre en place, elle nécessite également de structurer les données. Il devient alors beaucoup plus facile de maîtriser les accès en lecture, voire dans certains cas en écriture.

Cette étude amont doit également être celle de la mise en place d'une gouvernance : diffuser des informations, même en interne, équivaut à prendre le risque de les retrouver en externe (malveillance, inattention...). Il est impensable (et illégal avec le nouveau règlement RGPD), de libérer des données stratégiques ou personnelles. La question de la qualité doit également être abordée. Dans le cadre d'une diffusion à des fins de pilotage, un nettoyage sera nécessaire. Par exemple, la Haute Autorité pour la Transparence de la Vie Publique (HATVP) diffuse ses données sous forme de PDF, les rendant pratiquement inexploitables.

Enfin, la mise à disposition des données doit être progressive, afin de permettre à l'organisation de s'adapter aux changements de culture et de processus que cela entraîne.

**Le déroulé type d'un projet**

Dans un premier temps, il convient de libérer ses données en interne, afin de commencer à propager une culture centrée sur l'utilisation et la qualité. C'est également l'occasion de débuter des «Proof of concept». Cette étape ne saurait constituer, en aucun cas, un test pour déterminer ce qui est stratégique et ce qui ne l'est pas. La gouvernance doit être mise en amont comme nous avons pu le voir précédemment.

Dans un deuxième temps, les données peuvent être laissées à des partenaires externes, dans le cadre de projets communs. Cette étape permet de s'assurer que les données sont compréhensibles et utilisables par des 'extérieurs'. Ce sera alors le moment d'effectuer certains ajustements : métadonnées, accessibilité des portails, performances...

C'est seulement dans un troisième temps que les données pourront être libérées sous la forme d'Open Data au public.

Ces trois étapes imposent une bonne compréhension des besoins. Chacune d'elles doit avoir été individuellement menée à bien : l'objectif est d'avoir identifié les projets à fort ROI que l'on pourrait mener en interne avant de rendre les données publiques. Un projet d'ouverture doit donc être transverse afin de ne laisser passer aucune opportunité.

" La mise à disposition des données doit être progressive, afin de permettre à l'organisation de s'adapter aux changements de culture et de processus que cela entraîne. "

Pour conclure, mettre à disposition ses données peut être un véritable projet stratégique. Plus que tout, c'est l'occasion de mettre en place une culture de la donnée. En formant ses salariés à l'importance de la qualité et à la prise de décision basée sur des faits, les bénéfices peuvent être très importants. L'avantage compétitif obtenu par les géants d'internet, est essentiellement basé sur leur capacité à utiliser tout ce qu'ils collectent. Pour des entreprises plus « classiques », la difficulté réside dans l'initiation de ces processus en motivant les équipes.

# Brèves

## RETROUVEZ MC²I GROUPE AU CONGRÈS BIG DATA PARIS 2019

Comme chaque année, mc²i Groupe sera présent au congrès du Big Data qui aura lieu les 11 et 12 mars 2019 au Palais des Congrès de Paris. Pour sa huitième année, alors que le Big Data est au cœur des préoccupations des entreprises, ce sera l'occasion pour mc²i Groupe de vous rencontrer sur son stand et d'échanger avec de nombreux partenaires et start-up sur l'avenir de la donnée.


## RGPD : UNE TABLE RONDE SUR LES ENJEUX CONCRETS APRÈS LE 25 MAI

mc²i Groupe a convié ses clients lors d'un petit déjeuner pour partager avec eux des retours d'expérience sur le RGPD. Au travers de témoignages sur la mise en œuvre opérationnelle, de nombreux échanges ont eu lieu sur des problématiques variées telles que la gestion des consentements, les durées de conservation, les opportunités à saisir ou encore la conduite du changement.

## DES EXPERTS MC²I GROUPE AUX CÔTÉS DES ÉTUDIANTS

mc²i Groupe intervient chaque année dans plusieurs cursus d'enseignement supérieur, pour dispenser des cours de Data Analytics et Business Intelligence sous différents formats : Conférences, études de cas, jeux de rôles, exercices et soutenances.
Nos prochaines interventions sur la fin d'année 2018 sont prévues à l'Ecole Centrale Paris (ECP), au Master MIAGE à l'université de Paris Sud Orsay, à l'Université de Technologie de Troyes (UTT) et à l'EPF.

mc²i news LA LETTRE D'INFORMATIONS GROUPE

HORS SÉRIE NUMERO 9 - Octobre 2018

Directeurs de publication : Alexis TRENTESAUX, Vincent PASCAL, Marion LOPEZ
Rédacteurs en chef : Alianor SIBAI, Julie LAUV, Cédric BELHARRAT, Alexis MONNEROT DUMAINE, Aude LAMODIERE
Coordination : Carla LAMANNA
Rédaction : Alianor SIBAÏ, Perrine MAIRRE, Julie LAUV, Aurélien DEMACHY, Alexis GIGLEUX, Fabien LORENZI, Aboubacar BESSENGA DIALLO, Christophe DE BOISSET
Illustrations : Jérémy VARIN
Version digitale : Yann GOUPIL
Conception : Orkidées / Réalisation : J'magine
Crédits photographiques : iStock / Fotolia

# LE CONSEIL EN TOUTE INDÉPENDANCE

mc²i Groupe conseille et accompagne les grands groupes privés et les administrations dans les enjeux de la transformation numérique, autour de deux axes principaux : l'alignement du Système d'Information sur la stratégie de l'entreprise et l'adaptation des Organisations aux nouveaux modes de travail.


UN ACTEUR INDÉPENDANT

L'expertise et la valeur ajoutée des prestations de mc²i Groupe sont renforcées par son indépendance à l'égard des acteurs du marché, en particulier des éditeurs de solutions, des intégrateurs de systèmes et des infogéreurs.

Cette indépendance garantit l'intégrité des recommandations formulées aux clients, enrichies par une veille permanente et des benchmarks réguliers des solutions du marché.

51 rue François 1er - 75008 PARIS
Tél : 01 44 43 01 00
contact@mc2i.fr
www.mc2i.fr
Suivez-nous sur @mc2iGroupe