

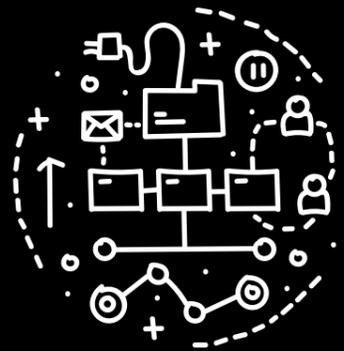
2017 / 2018

# GUIDE DU BIG DATA

L'ANNUAIRE DE RÉFÉRENCE  
À DESTINATION DES UTILISATEURS

BIG DATA





# Du **Big Data** à l'**analytics** aux **sciences** **cognitives**

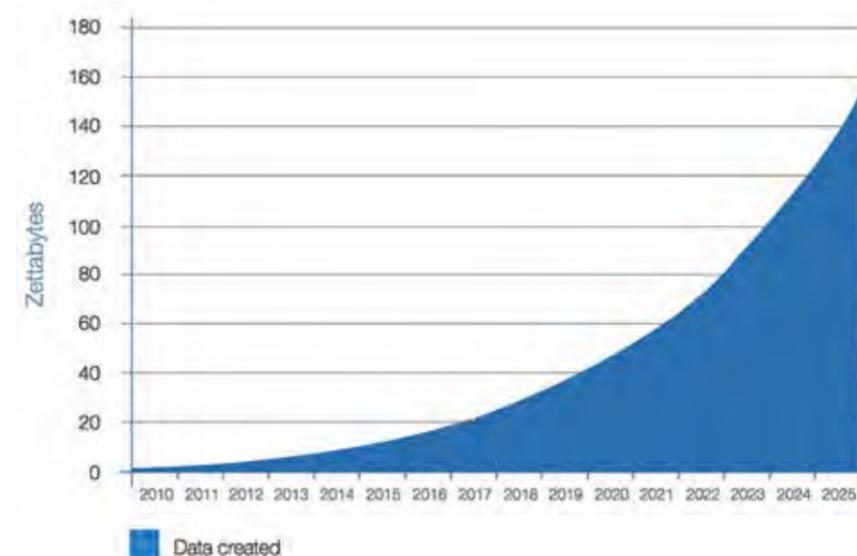


## Une **nouvelle** **ère** de la data

★ Nous vivons dans un Monde dominé par la data et les algorithmes. Selon une étude réalisée par la CNIL et l'IFOP, **80% des citoyens français sont conscients de la présence des algorithmes dans notre société**. Une société data-driven dans laquelle l'information est devenue une véritable devise monétaire. Les consommateurs utilisent quotidiennement des services gratuits des GAFAM en échange des données laissées sur leur passage et de l'autorisation donnée à ces géants du web de monétiser leur comportement sur Internet. Gartner prédit d'ailleurs que **d'ici 2021, 20% de toutes les activités individuelles impliqueront au moins un des sept géants de l'IT**, parmi lesquelles Google, Apple, Facebook, Amazon pour les États-Unis et l'Europe et Baidu, Alibaba et Tencent pour la Chine, générant toujours plus de données.

Plus globalement, l'IDC prédit que les **données produites en 2025 atteindront 163 Zettaoctets** (1 Zo équivalent 1 180 milliards de milliards d'octets), **soit 10 fois plus que les 16,1 Zo de données générées en 2016**. Face à cette augmentation considérable du volume de données, la question primordiale qui se pose n'est plus celle de la génération et collecte des données, mais bien celle de la qualité, de la **valorisation des données et des perspectives business qui peuvent en être retirées**.

Annual Size of the Global Datasphere



Source : IDC's Data Age 2025 study, sponsored by Seagate, April 2017

D'autant que l'essentiel des **datas générées, (plus de 60% selon IDC) le seront par les entreprises**, et proviendront de sources multiples telles que les systèmes embarqués et l'IoT, les systèmes cognitifs et d'Intelligence Artificielle, le mobile et les applications en temps réel etc...

Indépendamment de l'origine des données, les entreprises seront amenées à gérer plus de 97% de la « Datasphere » dans son ensemble.

Le mot Big Data n'est d'ailleurs plus totalement d'actualité, tant le volume est une problématique révolue. Gartner, lors de l'une de ses conférences aux États-Unis, a proclamé la mort de la Business Intelligence et l'avènement de la Data Analytics. Alors que la Business Intelligence consiste à examiner des sets de données collectées afin de tirer des conclusions sur les informations qu'ils contiennent et de prendre des décisions business plus éclairées, l'Analytics est une véritable data science. Les outils de Data Analytics sont déployés afin de prédire des comportements dans le futur et de construire des modèles prédictifs d'analyse.

**163 Zettaoctets,**

**c'est le volume estimé des**

**données produites en 2025,**

**soit 10 fois plus que les**

**16,1 Zo de données**

**générées en 2016**

## Trop de **data** !

★ Devant cette abondance de data, les opportunités sont nombreuses pour tirer profit du large volume de données et pourtant les entreprises n'ont pas encore les capacités et compétences nécessaires pour donner du sens aux données critiques, soit car la stratégie n'est pas adaptée, soit par manque de ressources ou de technologies. Selon un sondage réalisé par NewVantage Partners auprès des « Fortune 1000 entreprises », elles peinent encore à établir une culture data-driven : **69,4% d'entre elles ont lancé des initiatives en la matière mais seul 27,9% en voit les bénéfices. De même, plus de 45,2% d'entre elles n'ont pas encore commencé à monétiser leurs initiatives Big Data.**

Parmi les obstacles prioritaires rencontrés selon ce même sondage, **le manque d'alignement** entre les différents services au sein de l'organisation et la **difficile adoption et compréhension du Big Data par le middle management** sont cités en premier.

Le cabinet Gartner souligne ici **le rôle du Chief Data Officer**, une fonction encore peu développée en France comparé aux

États-Unis. L'intégration d'un Chief Data Officer (CDO) est pourtant **un indicateur clef qu'une organisation s'inscrit dans une culture data-driven.**

Le rôle du CDO consiste justement à rendre accessibles et compréhensibles les données par l'ensemble de l'entreprise, depuis la logistique jusqu'à la vente, en passant par les RH, le marketing, etc. Son rôle est donc lié à la gouvernance de ces informations, en relation avec les principaux décideurs de l'entreprise. La majorité des cadres interrogés par NewVantage estime d'ailleurs que les CDOs devraient reporter au CEO (53,4%) ou au COO (17,8%) plutôt qu'au CIO (15,6%). La relation entre le CIO et le CDO est souvent complexe, chacun revendiquant le périmètre de l'autre.

De même seulement 15% de l'ensemble des données sera taggué, et pourra donc être soumis à une analyse Machine Learning ou d'Intelligence Artificielle, et seulement 3% du total sera analysé par des systèmes cognitifs.

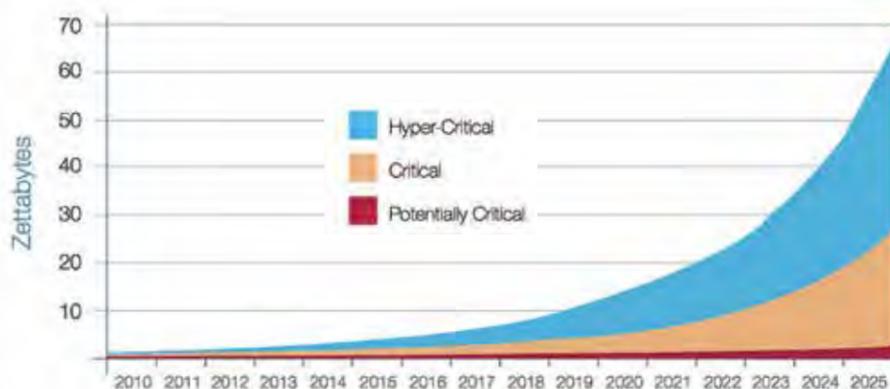
## Des données structurées aux données non structurées

★ Les données structurées sont celles dont l'ensemble des valeurs possibles est déterminé et connu à l'avance. Par exemple, dans une base de données rassemblant les résultats d'une enquête d'opinion, l'âge ou la catégorie socio-professionnelle des individus interrogés sont des données structurées, car les tranches d'âges ou la liste des catégories socio-professionnelles possibles sont déterminées a priori.

## Quelles datas ?

★ Naturellement, toutes les données ne sont pas adaptées ni disponibles à l'analyse. L'IDC estime ainsi que d'ici 2025, seulement **20% de la Datasphere sera essentielle** (« critical ») et seulement **10% sera absolument essentielle** (« hyper-Critical ») dans notre vie quotidienne.

### Data Critical Over Time



Data Type	CAGR 2015 to 2025
All Data. Includes all data in the global datasphere.	30%
Potentially critical. Data that may be necessary for the continued, convenient operation of users' daily lives.	37%
Critical. Data known to be necessary for the expected continuity of users' daily lives.	39%
Hypercritical. Data with direct and immediate impact on the health and well-being of users. (Examples include commercial air travel, medical applications, control systems, and telemetry. This category is heavy in metadata and data from embedded systems.)	54%

Source : IDC's Data Age 2025 study, sponsored by Seagate, April 2017

**90% des données sont**

**soit semi-structurées**

**soit non structurées**

Les données non structurées sont de façon générale des données textuelles qui ne rentrent pas dans des bases de données relationnelles. Elles existent sous de nombreuses formes (textes, images, vidéos, sons...), sont publiées sur de nombreux supports (articles de blog, fils Twitter, posts Facebook, réclamations téléphoniques auprès du service client...) et sont essentielles à la compréhension du business. **Alors que les données structurées décrivent une action ou un fait, les données non structurées donnent les raisons et le pourquoi !**

On estime aujourd'hui que **plus de 90% des données sont soit semi-structurées soit non structurées.** Jusqu'ici la valeur de ces données était largement sous exploitée et la priorité des entreprises était de mettre en place des initiatives sur les données structurées notamment parce que les outils nécessaires pour exploiter le volume des données non structurées n'existaient pas encore. **Les technologies de Machine Learning et de data visualisation rendent cela possible, de même que la baisse des coûts des technologies et la qualité de celles-ci** vont permettre aux entreprises de mettre l'analyse des données non structurées en haut de leurs priorités.

Par leur variété et leur volume, les données non structurées sont essentielles, par exemple pour améliorer la connaissance client. L'analyse linguistique et le text mining permettent désormais de reconnaître les mots et expressions soigneusement choisis pour décrypter les sentiments exprimés par l'internaute, le client ou le prospect.

## Des plateformes analytiques à l'analyse intégrée

★ L'analyse intégrée ou embarquée (embedded analytics) représente l'utilisation



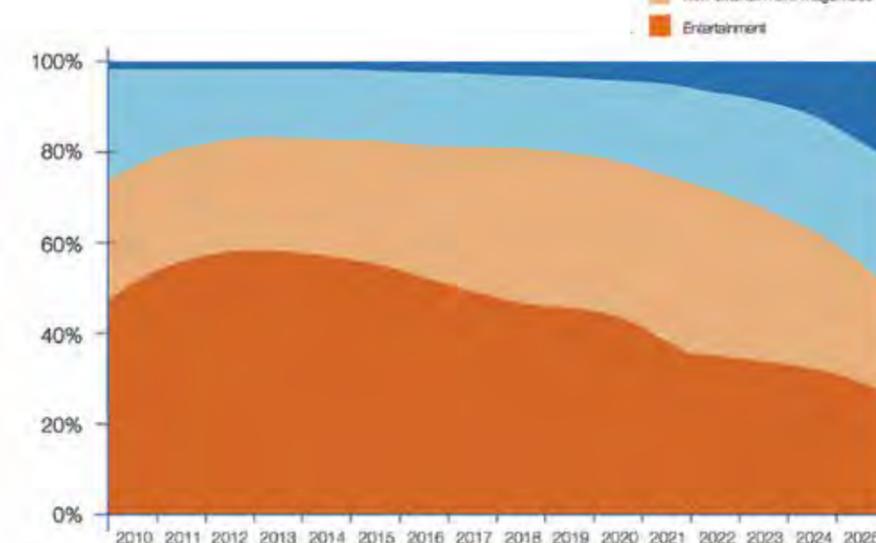
des capacités d'analyse au cœur même des applications logicielles. Les outils analytiques sont des composants natifs de l'application plutôt que des plateformes dédiées et distinctes de l'application. Ces capacités d'analyse peuvent résider en dehors de l'application, néanmoins elles doivent être accessibles de manière aisée et rapide par l'application, sans forcer l'utilisateur à passer par un autre système. Selon une étude Logi Analytics, **plus de 66% des équipes IT utilisent désormais l'analyse embarquée dans leurs organisations, et presque 30% envisagent de l'utiliser dans un futur proche.**

Les données embarquées sont produites par les objets connectés et les technologies Machine-to-Machine tels que les caméras de sécurité, compteurs intelligents, lecteurs RFID, wearables, les automates, les voitures & tous les moyens de transports connectés, les bâtiments connectés, implants médicaux etc. Selon Gartner, **26 milliards d'objets seront connectés dans le monde en 2020 et 6 milliards pourraient bénéficier d'Intelligence Artificielle dès 2018.**

Aujourd'hui, chacun d'entre nous détient en moyenne moins d'un objet connecté embarquant des capacités analytiques. Dans les 10 prochaines années, chaque personne en détiendra plus de 4 (Source : IDC) et les données des systèmes d'analyse embarqués représenteront presque 20% de l'ensemble de la « Datasphere » d'ici 2025.

Cependant, les données IoT ne provoqueront qu'une très légère augmentation de la demande de stockage (moins de 3%), d'ici 2018. La plupart des données IoT ne seront pas conservées.

### Data Creation Share by Type



Source : IDC's Data Age 2025 study, sponsored by Seagate, April 2017

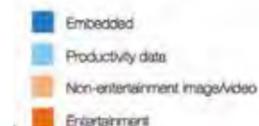
Les données collectées de l'IoT et des systèmes embarqués sont de vrais gisements de valeur, pourtant, selon Forrester, seul un tiers des entreprises qui exploitent des équipements connectés savent utiliser les données générées. Il faut donc **instaurer un dialogue entre data-scientists et responsables métiers afin de faire parler les données collectées.**

L'essor des objets connectés et embarqués impliquera naturellement une explosion du volume de données mobiles et de données en real-time va augmenter une fois et demi plus vite que le volume des données globales.

## Du prédictif au prescriptif !

★ L'analytique pour les entreprises est passée par plusieurs phases, qui ont abouti aux modèles suivants :

- **L'Analytique descriptive** qui aide les utilisateurs à avoir une meilleure compréhension des données passées grâce à du reporting, du groupement et des tableaux de bord.
- **L'Analytique prédictive** qui utilise des techniques statistiques pour étudier les données passées et présentes afin de faire des prévisions.
- **L'Analytique prescriptive** qui permet aux utilisateurs de prendre les bonnes décisions par la simulation et l'optimisation des scénarios.



En 2017, de plus en plus d'entreprises se sont intéressées à l'analyse prescriptive mais seulement 10% des organisations dans le monde l'utilisent, un chiffre qui devrait atteindre 35% d'ici 2020. Gartner prédit par ailleurs que le marché devrait atteindre 1,1 milliards d'euros d'ici 2019.

## Des technologies Big Data aux technologies cognitives et d'Intelligence Artificielle

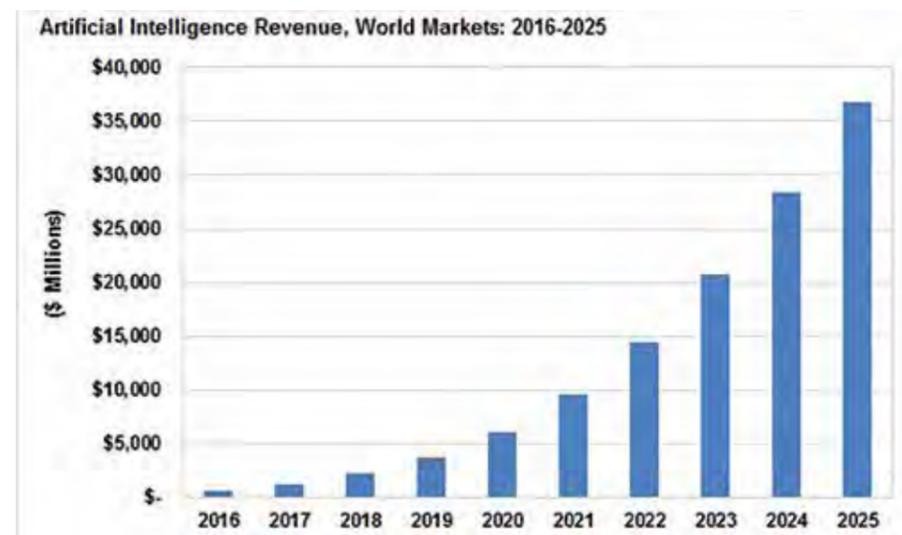
★ Alors qu'aujourd'hui, seulement 4% des organisations disposent d'un système cognitif opérationnel, les ¾ des organisations interrogées ont les data et les capacités analytiques requises pour implémenter des systèmes cognitifs. L'explosion du volume de données disponibles donne implicitement un avantage compétitif aux entreprises ayant intégré l'Intelligence Artificielle. 89% des early adopters des technologies cognitives sont plus rentables et plus innovants que leurs pairs et

concurrents. (Source : IBM)  
Outre les données nécessaires, les applications d'Intelligence Artificielle nécessitent des capacités de stockage en masse et une puissance de traitement que seules les toutes dernières technologies peuvent offrir. L'économie du Cloud jouera un rôle de premier plan dans le développement d'applications basées sur l'Intelligence Artificielle dans l'entreprise. Ainsi d'ici 2020, 67% des dépenses en infrastructures des entreprises IT seront basées sur des offres Cloud et les plateformes Cloud collaboratives utilisées conjointement par les entreprises vont tripler d'ici 2018 (Source : IDC).

Plus globalement, d'ici 2019, 40% des initiatives de transformation digitale et 100% des développements IoT seront supportés par les technologies cognitives et l'Intelligence Artificielle (Source IDC).

Le marché de l'Intelligence Artificielle pour les applications en entreprise est estimé à plus de 11 Milliards de dollars en 2024 et le revenu global est estimé à plus de 36 Milliards de dollars d'ici 2025 selon Tractica.

Une étude Accenture montre que l'Intelligence Artificielle a le potentiel d'augmenter le taux de profitabilité de 38% en moyenne et d'entraîner une relance de l'économie à hauteur de 14 milliards de dollars à travers 16 industries d'ici 2035. Les domaines de l'information et de la communication, de la finance et du manufacturing profiteront prioritairement de l'impact de l'Intelligence Artificielle.



Source : TRACTICA

## De la défense à la sécurité by design

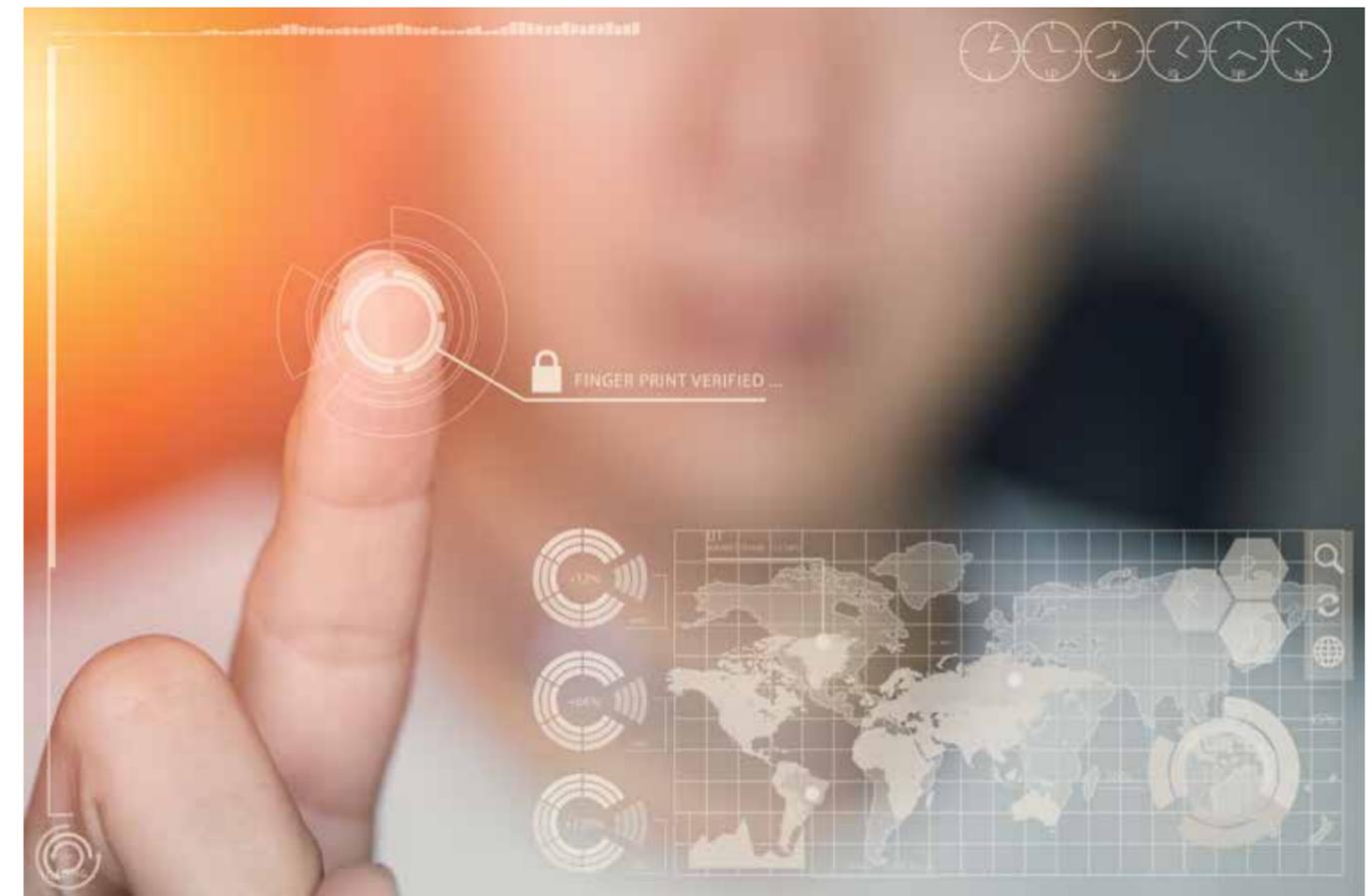
★ Avec de plus en plus de données personnelles à gérer, en real-time et mobiles, les entreprises se doivent d'investir massivement pour garantir la fiabilité et la confidentialité des données et manager les risques de sécurité. La plupart des données requièrent un haut niveau de sécurité, soit presque 90% d'entre elles d'ici 2025, à l'image des informations financières, des informations personnelles et médicales... D'autre part, une prise de conscience est nécessaire sur le fait que les challenges en termes de sécurité et de confidentialité ne sont pas uniquement dédiés aux équipes IT mais requièrent la participation de l'ensemble des salariés d'une entreprise.

Le Règlement européen sur la protection des données personnelles (RGPD) qui entrera en application le 25 mai 2018 vise notamment à renforcer les droits des personnes et responsabiliser les acteurs traitant des données (responsables de traitement et sous-traitants).

Nommer un référent Conformité Informatique & Libertés pour certifier les démarches, le former à ses missions de Data Protection Officer (DPO), former les collaborateurs à la collecte des données, faire des audits pour vérifier la conformité au regard des nouvelles directives, avoir le temps d'établir un plan d'actions, mettre en place des process et des plateformes SI "sécurisés" sont autant de recommandations de la CNIL pour être en conformité avec le règlement.

Néanmoins, la mise en conformité avec le règlement européen ne garantit pas pour autant la fiabilité, la transparence et l'éthique de l'ensemble des démarches Big Data et d'Intelligence Artificielle. Les algorithmes sont également remis en question. Selon une étude de la CNIL, plus de 72% des personnes interrogées en France estime que les algorithmes sont un véritable enjeu de société et plus de 64% considère les algorithmes comme une menace en raison de l'accumulation des données personnelles sur les choix, les goûts et les comportements de chacun. Il revient aux entreprises de s'associer aux institutions publiques pour instaurer la confiance des citoyens et consommateurs et définir des standards éthiques sur le sujet. TransAlgo interviewé ci-après intervient notamment dans ce but.

La plupart des données requièrent un haut niveau de sécurité, soit presque 90% d'entre elles d'ici 2025





## QUANTUM S'EMPARÉ D'UNE NOUVELLE NICHE DANS LE MARCHÉ DE L'ANALYTIQUE

En alliant une distribution MapR Hadoop à l'infrastructure Cisco UCS, le spécialiste de l'analytique diminue ses temps de requête de 92% et améliore la précision de ses résultats.

### LES BESOINS

Les consommateurs australiens sont parmi les plus avancés technologiquement. Valorisant la personnalisation de leur expérience, ils s'attendent à recevoir des messages, des recommandations et des promotions adaptées à chacun. De telles observations apportent une pierre à l'édifice de Quantum, société spécialisée dans la data analytique basée à Sydney. Depuis 15 ans, Quantum offre à ses clients un aperçu des besoins consommateurs, de leurs comportements, de leurs habitudes d'achats, des media qu'ils consultent en appliquant des techniques rigoureuses de data science aux données transactionnelles. Forte de son expérience dans les secteurs du retail, de la finance et des médias, l'entreprise travaille avec des sociétés telles que Woolworths, National Australia Bank ou Coca Cola. A l'origine, Quantum analysait uniquement les données fournies par ses clients. Un business très lucratif pour ses dirigeants, qui ont souhaité innover et acquérir leurs propres sets de données. L'objectif : offrir de nouveaux services à plus forte valeur ajoutée par le biais d'une corrélation entre les données internes de ses clients et des informations externes anonymisées sur le comportement d'achat des consommateurs.

Mais pour passer du concept au déploiement, Quantum a dû faire face à un obstacle de taille : la plate-forme analytique existante. Au fur et à mesure que la complexité et le périmètre de leurs analyses progressaient, la plateforme de base de données existante et l'architecture serveurs ne pouvaient plus faire face à la demande. Quantum a donc eu besoin d'une nouvelle plate-forme analytique reposant sur des serveurs fiables, évolutifs, et hautement performants, ainsi que sur un framework logiciel de classe entreprise.

### LA SOLUTION : UNE DISTRIBUTION HADOOP MAPR SUR UNE INFRASTRUCTURE CISCO

Pour relever le défi, Quantum a constitué un groupe de data-scientists afin de créer leur propre cahier des charges et d'évaluer les solutions logicielles et matérielles du marché. «Les décisions concernant la nouvelle plate-forme affecteront nos activités pour les années à venir. Nous avons donc beaucoup investi dans le processus de sélection», commente Alex Shaw, Responsable des opérations technologiques. Quantum s'est rendu compte qu'une solution Big Data était nécessaire non seulement en raison du volume de données, mais surtout pour répondre aux lourdes exigences d'analytique. Après s'être décidé sur Hadoop pour la partie logicielle, s'est posée la question de la distribution. Après un POC de trois mois durant lequel différents use-cases ont été testés sur des sets de données et de requêtes réelles, notre partenaire MapR, qui bénéficie du label « Cisco Preferred Solution », a remporté la mise. Niveau matériel, Quantum a souhaité short-lister son fournisseur historique, les équipements Cisco, et un fournisseur OEM de serveurs x86 dit « de commodité ». Si les résultats en matière de performance ont été relativement similaires, les serveurs Cisco UCS ont fait la différence tant sur leur faible consommation énergétique que sur leurs capacités de gestion centralisée.

Le portefeuille de Quantum se développant rapidement, l'entreprise a dû augmenter sa capacité de calcul en conséquence. L'architecture en cluster de la nouvelle plate-forme lui permet d'évoluer facilement, tout en conservant un TCO inférieur aux architectures précédentes : Quantum pouvant affiner son niveau de performance par un simple ajout de serveurs au cluster, sans avoir besoin de les remplacer par des modèles plus puissants. Comparativement, les outils de gestion de la plateforme Cisco/ MapR réduisent jusqu'à 30% les délais d'administration d'un système multi-tenant, où les données provenant de plusieurs clients résident sur un même cluster. Si l'approche « multi-tenant » réduit les coûts et optimise l'utilisation du matériel, elle introduit un risque d'accès non autorisé à l'information par d'autres utilisateurs du même cluster. « Pour garantir la sécurité des données de nos clients, nous nous reposons sur les fonctionnalités telles que le partitionnement de données qui permet de séparer logiquement les sets de données tout en optimisant le stockage », précise Alex Shaw.

### DES PERFORMANCES QUI DÉPASSENT LES OBJECTIFS DÉFINIS

Au cours de la rédaction du cahier des charges, Quantum s'est fixé comme objectif une augmentation de ses performances par un facteur dix. Cet objectif a été largement dépassé puisque la plateforme Cisco/MapR affiche une diminution du temps de traitement des requêtes de 92%, soit une progression en termes de performances par un facteur 12,5. La performance est cruciale pour Quantum qui souhaite fournir des résultats complets et rapides à ses clients. «Le fait d'accéder à des sets de données externes, que nous combinons aux données apportées par nos clients, nous confère une position unique sur le marché ». Alex Shaw détaille : « Une enseigne retail peut disposer d'une image assez bonne du comportement de ses clients dans ses propres magasins, mais connaît-elle la fréquence à laquelle ses clients achètent chez un concurrent ? En faisant se rejoindre des données concernant la consommation en magasin ou en ligne, avec des habitudes de consommation en matière de media, nous sommes capables de rendre les campagnes publicitaires plus efficaces. »

A mesure que les sets de données ont augmenté et que l'analytique s'est faite plus complexe, les méthodes d'échantillonnage sont devenues plus longues et compliquées. Mais grâce à la nouvelle solution Cisco/MapR, les data scientists peuvent dorénavant concevoir des requêtes complexes, les exécuter sur des volumes de plusieurs Terabytes, et obtenir des résultats plus précis en quelques minutes à peine. En permettant de tester des scénarios alternatifs rapidement et avec précision, et en raccourcissant les délais de développement, la plateforme contribue à favoriser l'innovation. « Nous avons beaucoup de gens intelligents mais qui ont été handicapés par la technologie et son incapacité à accompagner et à mettre en œuvre leurs idées. Maintenant qu'ils ont amélioré la façon de procéder à nos analyses, nous ouvrons la voie à la création de services nouveaux et innovants » conclue Alex Shaw.

Pour en savoir plus

- Les solutions Cisco pour Big Data, visitez : [cisco.fr/big-data](http://cisco.fr/big-data)



# Vos données sont partout. Augmentez la valeur métier de vos données.



Source IDC 2016



Retrouver les résultats de l'étude IDC :





## COYOTE INFORME SES 5 MILLIONS D'UTILISATEURS EN TEMPS RÉEL GRÂCE À LA TECHNOLOGIE COUCHBASE MOBILE

Coyote, assistant d'aide à la conduite communautaire, a fait appel à Couchbase, éditeur de logiciel spécialisé dans la technologie de base de données, pour informer du trafic ses 5 millions d'utilisateurs en temps réel.



Le système Coyote avertit plus de 5 millions d'utilisateurs des limitations de vitesse, des situations de dangers, des alertes climatiques ou encore des perturbations routières à travers de multiples plateformes (dispositifs connectés, applications, systèmes embarqués).

Ce sont près de 13 millions d'alertes qui sont remontées quotidiennement par le système avant d'être partagées aux millions d'utilisateurs. Ces données doivent être synchronisées, traitées et diffusées en temps réel afin d'offrir un service premium à cette clientèle exigeante.

Couchbase Mobile intervient dans la récolte, le traitement et la synchronisation de ces données que ce soit en mode On ou Offline. Sans perte de temps à rechercher du réseau, les alertes sont instantanément envoyées sur les milliers de devices permettant ainsi d'optimiser l'expérience des utilisateurs avec une solution complète combinant agilité, gain de temps et efficacité.



« Là où nous passions des heures à synchroniser les données depuis le serveur sur les plateformes, avec Couchbase, d'une simple manipulation, les données sont déployées instantanément sur les millions de devices !

Grâce à Couchbase, nous pouvons aujourd'hui consacrer plus de temps au développement de nouveaux services et au déploiement de notre outil à l'échelle mondiale.

Couchbase n'est pas seulement un fournisseur mais c'est surtout un partenaire » se félicite **Mathieu Brun Bellut, CIO de Coyote**

Avec l'utilisation de Couchbase Mobile, Coyote se différencie en proposant des services personnalisés. En effet depuis l'intégration de cette solution, Coyote a développé l'outil de data visualization 'Mes Stats' qui permet aux utilisateurs de prendre connaissance de leurs données (nombre de kilomètres parcourus, vitesse moyenne, temps passé dans les bouchons, etc.).

### À PROPOS DE COUCHBASE

Couchbase propose des technologies de données NoSQL à des entreprises axées sur l'innovation numérique. Dans le monde entier, des organisations choisissent Couchbase pour la flexibilité de son modèle de données, la facilité de son redimensionnement, ses performances et sa haute disponibilité pour construire des applications web, mobiles, et IoT d'entreprise. La plate-forme Couchbase inclut Couchbase Server, Couchbase Lite, première base de données NoSQL mobile, et Couchbase Sync Gateway. Couchbase est conçu pour réaliser des déploiements mondiaux, avec réplique configurable entre datacenters afin de renforcer la proximité géographique et la disponibilité des données. Tous les produits Couchbase sont des projets open source.

Couchbase compte parmi ses clients des leaders de l'industrie tels qu'AOL, Amadeus, AT&T, Cisco, Comcast, Concur, Disney, eBay, General Electric, Marriott, Nordstorm, Neiman Marcus, PayPal, Ryanair, Rakuten/Viber, Tesco, Verizon, Wells Fargo, ainsi que des centaines d'autres grandes entreprises.

Accel Partners, Adams Street Partners, Ignition Partners, Mayfield Fund, North Bridge Venture Partners et WestSummit Capital font partie des investisseurs Couchbase.

### Contact presse

Agence Point Virgule - Khéloudja SADAoui  
+33 (0)6 52 49 85 80 / +33 (0)1 73 79 50 72  
ksadaoui@pointvirgule.com

# DE L'EXPÉRIENCE CLIENT À L'EUPHORIE CLIENT

Bienvenue sur la première *Engagement Database* au monde



Consultez [couchbase.com](http://couchbase.com) dès aujourd'hui pour en savoir plus.

START A REVOLUTION

# MAPR TECHNOLOGIES SAS

NICOLAS POULAKAKIS

Regional Sales Director

+33(0)6 87 98 11 90

sales-southeu@mapr.com

47 Boulevard Georges Clémenceau

92415 Courbevoie France

www.mapr.com



MAPR  
DATA TECHNOLOGIES

## HISTORIQUE

Fondée en 2009 à San José en Californie, MapR Technologies est un éditeur logiciel visionnaire qui a créé une plate-forme unique permettant de gérer n'importe quel type de données à travers l'ensemble des Clouds (privés, publics ou hybrides).

Nous permettons à nos clients d'exploiter la pleine puissance de leurs données grâce au premier système de données nouvelle génération, la « MapR Converged Data Platform ».

## COEUR DE METIER

Avec la Converged Data Platform de MapR, les entreprises peuvent effectuer de l'analytique et des applications simultanément au fur et à mesure que les données sont générées afin de créer de nouvelles applications, plus modernes et plus intelligentes, pour surpasser leurs concurrents.

MapR aide ses clients à traiter des problématiques telles que les vues client 360°, la détection de fraude, la gestion de risque, les moteurs de personnalisation ou de recommandation, l'analyse prédictive ou les objets connectés, pour n'en citer que quelques-unes.

## SERVICES / PRODUITS PROPOSES

La Converged Data Platform de MapR permet le traitement immédiat et direct des fichiers, tables et streams sur la même plate-forme.

Contrairement aux environnements « inter-connectés » qui demandent l'intégration et l'orchestration complexe d'un grand nombre de technologies, la convergence offre une architecture simplifiée permettant une vision des données en temps réel, un environnement de sécurité unifié et cohérent à travers l'ensemble des composants, une meilleure utilisation des ressources et une charge d'administration réduite.

De plus, au-delà des services de support assurés en 24x7, MapR propose également des services de consulting ainsi que des formations en ligne ou sur site.

## SECTEURS CIBLES

Les marchés adressés par MapR sont essentiellement les services financiers (banque de détail, banque d'investissement, assurances), la grande distribution, l'industrie, les médias, télécoms, services publics, et bien évidemment le secteur des nouvelles technologies.

## COUVERTURE GEOGRAPHIQUE

MapR est présent dans plus de 20 pays à travers les 3 principaux continents (Etats-Unis, Europe et Asie-Pacifique). Cette présence permet à MapR d'être au plus près de ses clients, et ainsi de leur fournir les meilleurs services en terme de consulting, ainsi bien sûr qu'un support 24x7x365 en « follow the sun ».

## AVANTAGES CONCURRENTIELS

Historiquement, MapR est reconnu pour proposer la seule distribution Hadoop hautement disponible et dédiée aux applications critiques. En plus de cette proposition de valeur historique, MapR a intensifié ses efforts de recherche et développement pour réduire la complexité d'intégration des nombreux composants nécessaires à la mise en œuvre d'une plate-forme de gestion de données, tels que Hadoop, Spark, le NoSQL, et Streaming et les moteurs SQL. En faisant converger toutes ces technologies dans une unique plate-forme, MapR réalise un tour de force en simplifiant l'intégration de ces composants essentiels tout en garantissant la haute disponibilité et la sécurité de l'ensemble. En se positionnant bien au-delà de Hadoop, MapR offre à ses clients la seule plate-forme de gestion de données conçue pour supporter leurs projets les plus ambitieux et les plus critiques.

Les plus grandes entreprises, tous secteurs confondus, font confiance à MapR pour résoudre leurs plus importants défis et construire leur Data Platform de demain.

# ALL DATA ONE PLATFORM EVERY CLOUD

## LIMITLESS POSSIBILITIES

Learn more at  
[mapr.com](http://mapr.com)

Auchan

CISCO

CRÉDIT AGRICOLE

hp

NTTSecurity

Pierre Fabre  
Dermo-Cosmétique

SAP

xactly  
Incent right



Réussissez vos projets clients Big Data avec Data Platform as a Service, disponible en moins de 5 minutes !

## HISTORIQUE

Lancée en 2010 à l'initiative de Dassault Systèmes, Outscale fournit des services de Cloud Computing de Classe Entreprise, répondant aux exigences réglementaires et locales, aux entreprises souhaitant augmenter leur agilité business et déployer rapidement des business modèles à valeur ajoutée, en interne et pour leurs clients.

## CŒUR D'ACTIVITÉ

Présente en Europe, Amérique du Nord et Asie, Outscale met à disposition de ses clients un riche catalogue d'offres et services répondant aux problématiques Big Data des entreprises les plus agiles.

Outscale, propose le meilleur des technologies Big Data aux intégrateurs et délivre des plateformes prêtes à l'usage, pour lancer leurs projets Big Data en moins de 5 minutes.

Intégrant la plateforme de données convergente de MapR, le fournisseur de Cloud apporte une solution clé en main, flexible et économique pour accompagner les entreprises vers l'ère du Big Data. Du déploiement de simples PoCs au prototype et à la mise en production des applications, les plateformes dédiées au Big Data d'Outscale allient performance, résilience et haute disponibilité.

## DATA PLATFORM AS A SERVICE – SOLUTION BIG DATA À LA DEMANDE

Réussissez vos projets clients Big Data avec Data Platform as a Service disponible en moins de 5 minutes.

Pour traiter et valoriser de très gros volumes de données numériques de leurs clients, Outscale met à disposition des intégrateurs sa plateforme spécialisée dans le déploiement d'infrastructures de projets Big Data : **Data Platform as a Service**.

Avec Data Platform as a Service, basée sur les technologies Hadoop et MapR, Outscale permet aux intégrateurs de déployer leurs architectures Big Data les plus complexes sur des infrastructures souveraines. Flexible, élastique et performant, le Cloud Outscale est la solution clé en main pour les intégrateurs à la recherche d'un Cloud de classe entreprise et un accompagnement sur-mesure pour supporter le déploiement de leurs infrastructures de traitement de données.

Les données collectées et transformées sont hébergées en France dans notre solution de stockage optimisée pour garantir une haute performance et la sécurité de données.



L'offre Data Platform as a Service, basée sur MapR, permet aux intégrateurs d'accompagner leurs clients dans la création d'applications hautement distribuées et fortement scalables, grâce à la technologie MapR associée à la puissance du Cloud Outscale.

## Disponible en moins de 5 minutes !

Avec l'offre Data Platform as a Service, les intégrateurs démarrent leurs plateformes Big Data en choisissant la taille de cluster la mieux adaptée aux besoins de leurs clients.

En moins de 5 minutes, la plateforme MapR dédiée est opérationnelle et les ressources sont disponibles.

## MapR, un Hadoop production-ready !

Afin de permettre au marché d'avoir confiance dans le Big Data, MapR s'est appuyé sur ses équipes de R&D pour transformer Hadoop en une plateforme prête pour la production. Avec MapR-FS, MapR-DB et MapR-Streams, les projets Big Data ne sont plus cantonnés à des simples PoCs, ils deviennent le cœur des stratégies de développement et de digitalisation des plus grands acteurs.

## DES SOLUTIONS FLEXIBLES ADAPTÉES À TOUT TYPE DE BESOIN

Deux types de configuration de Data Platform as a Service sont disponibles pour choisir le cluster le mieux adapté à chaque projet Big Data :

### DataLab

- 1 noeud avec 16 cœurs et 64 Go de RAM
- Stockage utile : 100 Go

Plateforme dédiée aux Data Scientists pour affiner leurs algorithmes et parfaire leurs compétences.

### Enterprise

- 5 noeuds avec 16 cœurs et 64 Go de RAM par noeud
- Stockage utile : 2 To

Plateforme prête pour la production associant performance, haute-disponibilité et forte résilience.

## BÉNÉFICES DE DATA PLATFORM AS A SERVICE



**Facilité et rapidité de mise en place**  
Un cluster Hadoop correspondant à vos besoins en moins de 5 minutes



**Facturation de licence à l'heure**  
Un modèle de facturation agile vous donnant accès au support MapR



**Sécurité des données**  
Un Cloud Provider respectant la protection et la sécurité des données en France

En savoir plus sur l'offre Data Platform as a Service : [fr.outscale.com/dataplatform-as-a-service](http://fr.outscale.com/dataplatform-as-a-service)



**SERGE BOULET**  
Directeur marketing

01 60 62 11 11  
comsas@sas.com  
Domaine de Grégy - Grégy-sur-Yerres  
77257 Brie Comte Robert Cedex France  
www.sas.com/france



## SAS

### HELGE ROSEBØ

Architecte et Responsable Visualisation des Données

## INTERVIEW



INTERVIEW

### MUNICH RE MET EN PLACE UNE PLATE-FORME BIG DATA AVEC SAS ET HORTONWORKS

Le plus grand réassureur mondial, a retenu SAS® Analytics et Hortonworks Data Platform pour déployer une plate-forme innovante capable d'analyser des volumes gigantesques de données.

Pour Munich Re, la donnée est une composante essentielle de sa stratégie pour prendre en compte activement l'évolution mondiale du risque. Lancée en février 2016, la plate-forme permet aux différents départements d'explorer de nouvelles idées, de développer de nouveaux domaines commerciaux et d'améliorer le service à la clientèle.

L'analyse des données semi-structurées et non structurées, de données météorologiques ou de capteurs sur des véhicules, machines et autres dispositifs en réseau, font émerger des idées de produits, services et de nouveaux processus.

La plate-forme basée sur une approche ouverte fournit aux data scientists de Munich Re le plus haut degré de liberté et de créativité. Un lab leur permet de tester les idées sans affecter les opérations quotidiennes. Environ 500 utilisateurs travaillent déjà avec le système, qui est en cours d'extension par étapes.

### LE FEU A DÉCUPLÉ L'INTELLIGENCE HUMAINE. COMMENT S'EN INSPIRER ?

Nous consacrons une grande part de notre existence à faire des choix. En essayant de faire les meilleurs possibles. Baser les décisions sur des faits importe plus que le type et la qualité des données qui l'alimentent. En ancrant cette approche dans votre culture d'entreprise, vous saurez tirer parti du vaste potentiel qu'offrent les données. **Par Helge Rosebø, Architecte et responsable visualisation des données, SAS Norvège.**

La ressemblance entre cet état d'esprit et l'évolution de l'intelligence humaine me paraît frappante. En utilisant le feu comme technologie, nous avons décuplé notre intelligence. De quelle manière votre entreprise pourrait-elle développer la sienne ?

### Le feu nous a rendus plus intelligents

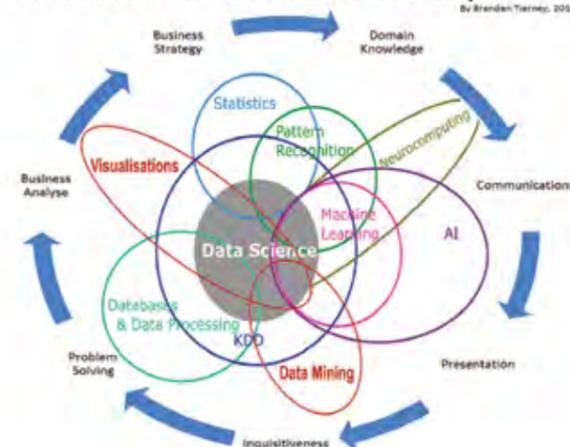
Le feu a eu un énorme impact sur le mode de vie de nos ancêtres, en leur apportant du confort et en leur permettant de cuire leurs aliments. La viande est plus digeste cuite que crue. L'organisme utilise donc moins d'énergie à assimiler un steak qu'un morceau de viande crue. Moins d'énergie pour digérer signifie également plus de vitalité pour d'autres fonctions corporelles. Autrement dit, la cuisson a contribué à développer notre intelligence, et le reste appartient à l'histoire. Nous sommes aujourd'hui arrivés à un stade où les nouvelles technologies pourraient bien surpasser nos **capacités cognitives**. Auparavant, le feu était la technologie charnière et la viande était une source d'énergie. La combinaison de la viande (données), du feu (technologie) et de la cuisson (processus) a dopé notre intelligence.

### DANS LE PAYSAGE NUMÉRIQUE ACTUEL, COMMENT DEVENIR PLUS INTELLIGENT QUE NOS CONCURRENTS ?

....Grâce à la science des données !

Les applications de la data science ont pour finalité de traduire les nouvelles connaissances en croissance économique. Le diagramme de Brendan Tierney offre un aperçu des compétences et du processus qui composent cette science. Il me sert d'aide-mémoire, et vous pouvez vous aussi l'utiliser pour replacer ce phénomène ultra-médiatisé dans le contexte de vos tâches et compétences actuelles.

### Data Science Is Multidisciplinary



S'il n'existe encore aucune technique d'intelligence artificielle ou de machine learning capable de faire disparaître tous nos problèmes d'un coup de baguette magique, elles sauront en revanche augmenter votre efficacité. Et celle-ci atteint son apogée lorsque l'on associe un panel de compétences pour répondre à des problématiques métier en constante évolution. Si l'on sait quelles compétences appliquer à une problématique donnée, c'est gagné !

### « Faire des concessions sur la technologie impacterait l'efficacité du processus. »

L'application de la science des données peut impliquer différentes fonctions (développeur, statisticien, analyste métier, expert en entretêts de données, chef de projet, etc.), ce qui rend l'accès à la réussite plus complexe. L'équipe de data science doit être réactive, souple et rapide, et ses travaux doivent pouvoir être aisément partagés à d'autres branches de l'entreprise. Elle a besoin d'une technologie qui valorise efficacement ses différentes contributions et compétences, et qui combinerait par exemple des solutions Open Source et propriétaires. Quoi qu'il en soit, les équipes ont besoin d'un environnement technique propice au bon déroulement du processus, du développement à la production, en passant par les tests et l'évaluation qualité. Faire des concessions sur la technologie impacterait l'efficacité du processus. Mais comme toujours, une excellente technologie ne peut que libérer le potentiel, le reste relève de notre responsabilité et de nos processus.

*L'intelligence artificielle comprend de nombreux sous-domaines :*

**L'apprentissage automatique (machine learning)** automatise le modèle analytique. Il utilise des méthodes à partir de réseaux de neurones, de statistiques, de recherche opérationnelle et de physique pour trouver des informations cachées dans les données, sans être programmé pour savoir où chercher et quoi conclure.

**Un réseau neuronal** est une sorte de machine learning inspiré du fonctionnement du cerveau humain. C'est un système informatique composé d'unités interconnectées (comme les neurones) qui traite l'information en répondant aux entrées externes, en relayant l'information entre chaque unité. Le processus nécessite plusieurs passages aux données pour trouver des connexions et dériver des données non définies.

**L'apprentissage profond (deep learning)** utilise d'énormes réseaux de neurones avec de nombreuses couches de traitement, profitant des progrès de la puissance de calcul et de l'amélioration des techniques pour l'apprentissage de modèles complexes sur de grandes quantités de données. Les applications courantes incluent la reconnaissance de l'image et de la parole.

**La vision par ordinateur** repose sur la reconnaissance des formes et l'apprentissage profond pour reconnaître ce qui se trouve dans une image ou une vidéo. Lorsque les machines peuvent traiter, analyser et comprendre des images, elles peuvent capturer des images ou des vidéos en temps réel et interpréter leur environnement.

**Le traitement du langage naturel** est la capacité des ordinateurs à analyser, à comprendre et à générer du langage humain. La prochaine étape de cette science est l'interaction du langage naturel, qui permet aux humains de communiquer avec les ordinateurs en utilisant une langue normale et quotidienne pour effectuer des tâches.

Struggling to develop and deploy data analytics?

## LET'S CHANGE THAT™



01 45 81 81 00  
masters@telecom-paristech.fr  
46 rue Barrault  
75634 Paris Cedex 13 France  
www.telecom-paristech.fr/bigdata



## LE BIG DATA, UNE EXPERTISE AU CŒUR DE TÉLÉCOM PARISTECH

Première grande école française d'ingénieurs dans le domaine des Technologies et des Usages du Numérique, Télécom ParisTech forme ses diplômés à innover et entreprendre dans un monde désormais numérique. Près de 1500 étudiants la choisissent chaque année dont 55 % d'internationaux. Toutes les disciplines fondamentales des Technologies et Usages du Numérique (y compris les Sciences Economiques et Sociales), sont couvertes par ses enseignements et sa recherche.

Télécom ParisTech a une forte prise sur le monde socio-économique et le pourvoir en ingénieurs, docteurs et professionnels, formés pour accompagner la transformation numérique des entreprises, de l'industrie et de la société. Télécom ParisTech transfère à l'industrie les résultats de sa recherche de très haut niveau, via des brevets, des licences et des start-ups, notamment au sein de son incubateur, ParisTech Entrepreneurs. École de l'IMT, membre fondateur de ParisTech et de l'Université Paris-Saclay, Télécom ParisTech se définit comme le Collège de l'innovation par le numérique de Paris-Saclay, dont l'ambition est de devenir l'un des premiers pôles d'innovation mondiaux.

## UNE RECHERCHE PLURIDISCIPLINAIRE UNIQUE

Portée par 170 enseignants-chercheurs permanents et 300 doctorants, sa recherche, de réputation internationale est évaluée A+ par l'AERES et est couronnée chaque année par des prix prestigieux (Médaille d'Argent CNRS, ERC, Google Awards, etc.). Ses laboratoires allient des expertises disciplinaires de très haut niveau dans les sciences et technologies de l'information et une capacité rare de mobilisation transversale interdisciplinaire.

Le dynamisme de la filière Big Data à Télécom ParisTech découle d'une recherche pluridisciplinaire autour des mathématiques appliquées, de l'informatique et des sciences humaines et sociales. Elle réunit une cinquantaine d'enseignants-chercheurs et autant de doctorants. En partenariat avec une quinzaine d'entreprises, l'école s'investit dans deux chaires de recherche dans le domaine du Big Data, ainsi qu'une chaire d'enseignement. La chaire Machine Learning for Big Data conduit ses recherches à l'interface des mathématiques et de l'informatique. Le machine learning vise à élaborer des algorithmes permettant aux machines d'apprendre automatiquement à partir des données et ainsi d'améliorer leurs performances. La chaire Valeurs et Politiques des Informations Personnelles traite des aspects juridiques, techniques, économiques et philosophiques qui concernent la collecte, l'utilisation et le partage des informations personnelles.

## UN ENVIRONNEMENT D'ENSEIGNEMENT EXCEPTIONNEL

Télécom ParisTech a bâti un écosystème unique, basé sur une forte interaction entre formation, recherche et innovation et qui se caractérise par l'accueil régulier de conférences et de colloques internationaux, un incubateur qui héberge plusieurs start-up issues du Big Data, des partenariats avec les entreprises clé du secteur... Cet environnement a permis au Mastère Spécialisé® « Big Data » d'être le premier du genre en France et de faire figure de référence dans le domaine. Par la suite, l'école a mis en place une large gamme de formations pour tous les besoins, notamment dans le cadre de Télécom Evolution. Télécom ParisTech entretient

des liens étroits avec l'industrie : une dizaine d'entreprises sont membres du comité de veille et de perfectionnement des formations et plusieurs dizaines participent aux enseignements, études de cas, mises en situation professionnelle, tables rondes et séminaires. Elles proposent également stages, thèses professionnelles, projets de groupe mais aussi données réelles anonymisées. En 2017, l'école a créé, en partenariat avec des entreprises, deux nouvelles chaires d'enseignement sur la science des données.

## PRIORITÉ À L'INNOVATION

L'incubateur ParisTech Entrepreneurs a accueilli en 15 ans plus de 300 start-up innovantes du numérique. Beaucoup d'entre elles utilisent les technologies du Big Data, que ce soit dans l'analyse et le traitement des données, le marketing et le e-commerce, l'industrie, les transports, la santé, la high-tech, les services Internet, la finance... L'entrepreneuriat est en effet fortement encouragé au sein des différentes filières de l'école. Télécom ParisTech participe également à la plateforme souveraine Teralab qui a pour but de lever la barrière de l'adoption des technologies Big Data. Infrastructure puissante, flexible et de confiance, Teralab accompagne ses utilisateurs sur les plans technologie et juridique. Elle héberge des projets portés par les entreprises, les start-up ou les organismes de recherche et de formation.

## TÉLÉCOM ÉVOLUTION, DES COMPÉTENCES TOUJOURS A LA POINTE

Télécom Evolution est le centre de formation continue commun à IMT Atlantique, Télécom ParisTech et Télécom SudParis. Il accompagne les entreprises et les organisations dans la montée en compétences de leurs collaborateurs autour des grandes problématiques du numérique : Internet des objets, Big Data, cybersécurité, transports intelligents... Télécom Evolution propose une gamme complète de formations depuis l'initiation jusqu'à l'expertise, dans une grande variété de modalités pédagogiques, du MOOC à la formation certifiante, dispensés par les meilleurs experts et des professionnels reconnus.



Les tables rondes entreprises font salle comble



Gilles Babinet, Digital Champion de la France auprès de la Commission Européenne

## MASTÈRE SPÉCIALISÉ® BIG DATA :

### GESTION ET ANALYSE DES DONNÉES MASSIVES



Un Bac+6 pour des diplômés en poursuite d'études ou en reconversion, accrédité par la Conférence des Grandes Écoles. Plus de 600 heures sur 9 mois de cours, travaux pratiques, séminaires, et un projet de groupe « fil rouge » proposé par une entreprise.

Un stage et une thèse professionnelle en entreprise sur 4 à 6 mois concluent un programme conçu et enseigné par des spécialistes du domaine, en forte adéquation avec les besoins des entreprises.

Le MS Big Data combine un large socle de connaissances techniques et une approche complète des aspects juridiques, économiques, business.

**91 % des diplômés** ont trouvé un travail facilement ou très facilement, et **75% avant la fin de leur stage\***.

\*Étude Télécom ParisTech, novembre 2016



## CES DATA SCIENTIST

### (CERTIFICAT D'ÉTUDES SPÉCIALISÉES)

**Pour les professionnels en poste** qui souhaitent une certification en science des données (collecte, stockage et calcul distribués, analyse statistique, visualisation).

**Très opérationnel**, le CES permet la maîtrise des techniques de gestion et d'analyse des Big Data et des principaux algorithmes du machine learning.

**12 sessions de 2 jours sur 10 mois** : cours, travaux dirigés, travaux pratiques, témoignages de professionnels et une plateforme d'e-learning.



## MOOC : FONDAMENTAUX POUR LE BIG DATA

**Gratuit, flexible et indispensable** pour valider une inscription au **Mastère Spécialisé®** ou au **CES**.

**En 6 semaines**, révisez les prérequis essentiels dans les domaines de l'analyse, de l'algèbre, des probabilités, des statistiques, de la programmation Python et des bases de données.

**7 modules**, des vidéos, des quizz et un projet final, et plus de **7 000** inscrits par session !



## FORMATIONS INTER-ENTREPRISES

- Big Data : enjeux stratégiques et défis technologiques
- Big Data : panorama des infrastructures et architectures distribuées
- Big Data : premiers succès et retours d'expérience
- Data Science : introduction au Machine Learning
- Data Science avec Python **[NOUVEAU]**
- Data Science avec R **[NOUVEAU]**
- Data Science dans le Cloud : Big Data, statistiques et Machine Learning
- Droit, RGPD et protection des données dans le big Data **[NOUVEAU]**
- Extraction d'informations du Web
- Opinion mining : e-reputation et recommandation **[NOUVEAU]**
- Text-Mining
- Visualisation d'information (InfoVis)



# LEAD The 4<sup>th</sup> DIMENSION

100 Speakers  
200 Marques  
15 000 Participants

