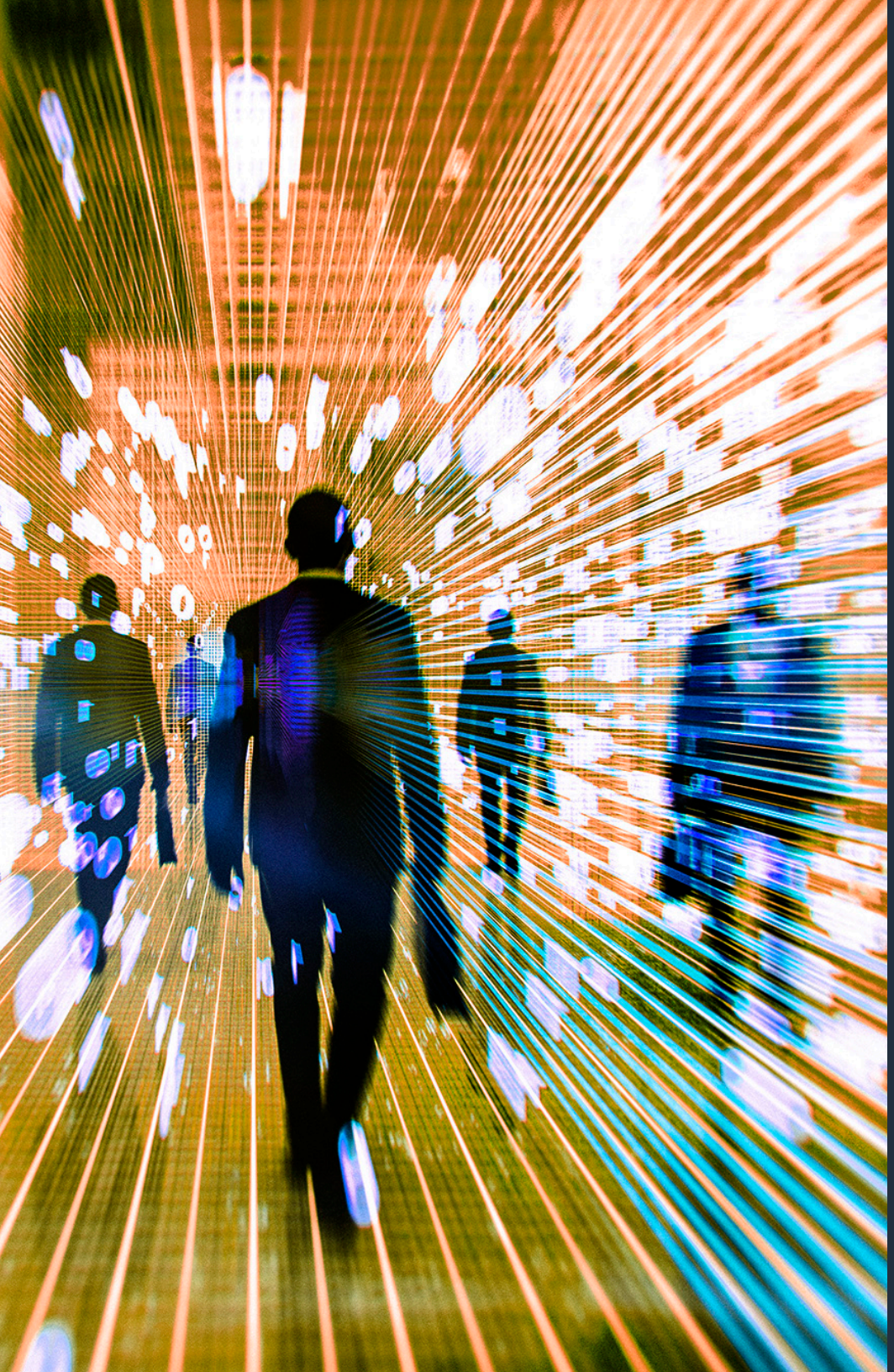
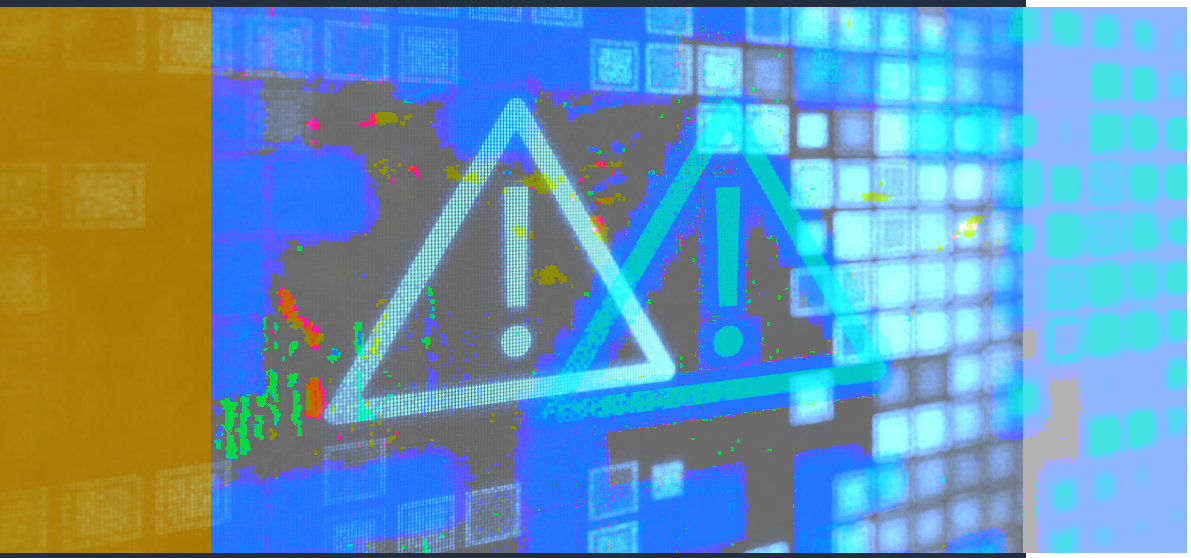


Devenir une entreprise pilotée par les données

Joe Chung, Stratège d'entreprise et évangéliste chez Amazon Web Services

Toutes les entreprises ont un problème de données



Imaginez ceci...

Le rapport Excel hebdomadaire est publié et il vous est envoyé directement par e-mail. En l'étudiant, vous constatez une anomalie dans les données financières que vous ne comprenez pas, et ce malgré le tableau croisé dynamique fourni dans le rapport qui vous permet d'accéder à un certain niveau de détail. Vous demandez des explications à votre analyste Opérations. Votre analyste vous répond : « Je ne sais pas. Je vais me renseigner ».

Le lendemain, l'analyste vous explique que cette anomalie est due à une baisse de la productivité à l'usine de fabrication.

> « C'est absurde, répliquez-vous. Pouvez-vous demander aux RH si les arrêts maladie ont un impact sur la productivité ? À moins qu'il y ait eu un problème avec l'application de contrôle de gestion de travail à l'usine » ?
>
> « Il va nous falloir une semaine pour obtenir ces données et les fusionner avec les données financières », répond votre analyste.
>
> « Envoyez-moi simplement une copie des données de l'ERP et de l'application de gestion du temps, je m'en occuperai moi-même ».
>
> L'analyste répond : « Je n'ai pas accès aux données, et il faudra quelques jours pour soumettre les bons tickets afin d'y avoir accès ».

Si ce scénario vous semble un peu trop familier, c'est que votre entreprise rencontre un problème de Big Data.

Votre première réaction sera peut-être de nous dire qu'il s'agit d'un problème d'outils et de processus de veille économique que connaissent les entreprises depuis la nuit des temps, pas d'un réel problème de Big Data. Sans entrer dans un débat dogmatique sur le rapport entre l'analyse, le reporting et la veille économique, le fait essentiel est que ***toutes* les entreprises ont un problème de Big Data.** Avec les capacités de l'intelligence artificielle et du machine learning qui commencent à se concrétiser, il est aujourd'hui plus important que jamais pour les entreprises de maîtriser les données qu'elles possèdent et de les exploiter au mieux pour devenir une **entreprise pilotée par les données**.

Le modèle mental que je propose pour décrire une entreprise pilotée par les données (Data-driven) est celle du système nerveux du corps humain. Les terminaisons nerveuses s'étendent dans tout le corps, envoyant à la moelle épinière des signaux sensoriels qui seront traités par le cerveau pour engendrer une réaction. Ce modèle est imité par des architectures de données capables de recevoir, de traiter et de stocker des données en temps réel, de l'intérieur et de l'extérieur de l'entreprise. Les signaux sont traités en temps réel et analysés par des algorithmes d'apprentissage automatique. Malheureusement, beaucoup trop d'entreprises estiment que ces nouvelles fonctionnalités sont loin d'être essentielles ou ne s'appliquent que dans des scénarios de données spécialisés ; certaines tentent même de faire passer leurs anciennes plateformes d'intelligence d'affaires pour des « lacs de données ».

Dysfonctionnements des données

Beaucoup d'entre nous s'imaginent que les problèmes de Big Data sont uniquement liés au volume. En réalité, chaque entreprise a des problèmes de données, au-delà du volume, qui ont été masqués de différentes manières. Voici quelques dysfonctionnements courants liés aux données que j'ai pu observer :

Des données isolées et éliminées

Tout d'abord, beaucoup d'entreprises ne se rendent pas compte que de nombreuses données exploitables sont effacées ou inaccessibles. Par exemple, des données telles que l'activité de l'utilisateur pour une application (et son utilisation de l'application en rapport avec d'autres applications) ; la télémétrie de l'infrastructure hébergeant l'application ; ou d'anciennes versions des données qui ne sont plus compatibles avec les schémas actuels.

Ensuite, les données sont réparties entre de nombreuses applications et entrepôts de données. Une seule application n'est peut-être pas très massive, mais l'ensemble de ces données l'est. Ainsi, lorsque l'entreprise a besoin d'analyser des données en provenance de sources multiples, cela devient très difficile. En effet, les données compartimentées posent un problème d'accès. Chaque emplacement de stockage possède ses propres règles et processus d'accès qui peuvent compliquer la gestion des données.

"

Chaque entreprise a des problèmes de données qui ont été masqués de différentes manières."

Données à basse fidélité

En règle générale, les anciens systèmes d'entreprise ne traitent et capturent que les états finaux, et leurs rapports ne concernent que de courtes périodes. De plus, les données sont traitées par lots et non en temps réel. Les données peuvent changer considérablement entre les fenêtres de traitement par lot, mais les anciens systèmes sont souvent conçus pour ignorer les changements transitoires, car ils ne sont pas en mesure de gérer la vitesse à laquelle peuvent évoluer les données.

Des données rondes dans des tables carrées

De nombreuses entreprises se rendent comptent qu'il existe des trésors de données qui ne sont pas adaptées aux technologies de stockage traditionnelles (images, données de capteurs, etc.). Il existe également mille et une manières d'analyser et d'exploiter les données. Par exemple, au lancement d'une nouvelle procédure analytique, vous pourriez vous rendre compte qu'aucune solution de reporting ou de visualisation ne peut répondre aux besoins de tous vos utilisateurs. La solution pourrait être de fournir des informations traitées par des algorithmes via des API, dans des applications utilisant des widgets de visualisation personnalisée utilisant des infrastructures JavaScript telles que D3.js, et via des portails d'intelligence d'affaires exploitant Tableau et autres solutions de visualisation.

Données désordonnées

Les systèmes d'entreprise n'aiment pas les données en désordre. C'est pourquoi il existe des formulaires, des règles et d'autres validations pour s'assurer que les données soient aussi nettes que possible avant leur stockage. Mais certaines données, parfois les plus intéressantes, ne sont pas aussi nettes. Quand on commence à puiser dans des données non structurées ou basées sur des objets, il y a forcément du bruit. Et tout comme avec le bruit électronique, vous pouvez vous servir de mécanismes de filtrage, d'amélioration et d'amplification afin d'exploiter ces données. Un cas d'utilisation qui préoccupe de nombreuses entreprises est la montée en flèche des coûts liés à l'envoi de données dans des outils propriétaires d'agrégation, de sécurité ou de suivi. Dans la plupart des cas, il est possible de réduire les coûts en filtrant une grande quantité de données de journaux n'étant pas utile.

Si vous vous êtes reconnu dans l'un des points ci-dessus, il est temps pour votre entreprise de repenser sa démarche et son architecture analytiques. Chaque entreprise a la possibilité de déployer des solutions analytiques adaptées à ses besoins (stockage, traitement, interrogation, analyse, présentation, etc.) afin de relever ses défis métiers et informatiques.

Les plateformes analytiques modernes offrent un insight métier crucial

Une fois prêt à résoudre votre problème de Big Data, que pouvez-vous raisonnablement attendre d'une plateforme analytique moderne ? Voici, d'un point de vue technique, ce qu'il est possible de réaliser aujourd'hui, et comment.

> "Les décisions basées sur les données nécessitent l'accès à de nombreux types d'informations disparates."

Accéder à toutes les données que je veux

Les décisions basées sur les données nécessitent l'accès à de nombreux types d'informations disparates. Un pilote s'appuie sur les jauges de l'avion pour comprendre les informations essentielles au vol, telles que l'altitude, la vitesse de l'air et la consommation de carburant. Mais imaginez que le pilote ne dispose pas de toutes ces jauges au même endroit. Peut-être doit-il se rendre en cabine, réclamer l'information par radio, ou pire encore, demander l'autorisation d'accès aux données. Malheureusement, il s'agit d'une réalité quotidienne dans le monde de l'entreprise d'aujourd'hui.

Des entreprises innovantes ont cependant renversé cette tendance en extrayant leurs données de leurs systèmes pour les stocker en un seul endroit (un lac de données). Bien que de nombreuses entreprises stockent de grandes quantités d'un seul type de données, de plus en plus d'entreprises créent des lacs de données à l'échelle de l'entreprise afin de stocker plusieurs types de données provenant de sources différentes.

Au début des années 2000, des entreprises à l'échelle d'Internet telles qu'Amazon, Yahoo et Facebook ont commencé à constater que les technologies de bases de données relationnelles avaient atteint leurs limites en termes de scalabilité et de performances. Amazon a réagi avec une technologie du nom de Dynamo, une base de données clé-valeur hautement disponible et scalable, comme la technologie NoSQL/non-relationnelle. Puis Amazon a évolué et exploité Dynamo pour créer des services tels qu'Amazon S3 et Amazon DynamoDB. Grâce à sa capacité à stocker de nombreux types de données et à son faible coût de stockage, Amazon S3 est intéressant pour les entreprises cherchant à créer des lacs de données. Il existe bien sûr d'autres solutions techniques, notamment Hadoop, mais une caractéristique importante de toutes les solutions de lacs de données est leur capacité à stocker tous les types de données à l'échelle du pétaoctet et à faible coût.

Réactivité au changement

Les systèmes et les données des entreprises évoluent en permanence, mais les systèmes chargés de rapporter ou de partager ces informations sont souvent les derniers à changer. Combien de fois vous a-t-on dit qu'il faudra six mois ou plus pour que les données soient restaurées dans les entrepôts et dans les rapports ? Ou que les nouvelles données du système source n'ont pas encore été transférées aux systèmes de reporting, et qu'il faut plusieurs jours pour que les modifications soient effectives en raison du traitement par lots ? **La vitesse à laquelle les données sont disponibles détermine la vitesse à laquelle les décisions peuvent être prises.** Nous devons donc attendre des systèmes analytiques modernes qu'ils soient capables de traiter et de rapporter les données en temps quasi réel, ainsi que d'être réactifs face aux modifications apportées aux sources de données en amont.

" La vitesse à laquelle les données sont disponibles détermine la vitesse à laquelle les décisions peuvent être prises."

Le premier facteur clé est la nature du stockage des données dans des technologies Big Data telles qu'Amazon S3 ou Hadoop. L'un des principaux obstacles à la modification d'une base de données relationnelle est la modification du schéma ou de la définition du mode de stockage des données. Tant que le schéma n'est pas modifié, les données ne peuvent pas atterrir dans la base de données sans être endommagées. Les technologies basées sur les fichiers ou les objets, comme Amazon S3, ne s'attachent pas à la structure des données. Les données peuvent venir telles quelles, par opposition à l'approche exigeant qu'elles s'adaptent à votre structure.

L'autre problème rencontré est que seul un schéma à la fois peut être actif. Bien sûr, nous avons tous vu des tables de base de données nommées « 2015 » et « 2016 », mais ce système est loin d'être idéal. Les technologies Big Data possèdent un schéma basé sur une approche en lecture, ce qui signifie que la structure des données est appliquée lorsque vous les récupérez, et non déduite en fonction de la façon dont elles sont stockées. Pour les entreprises, cela signifie que les modifications de données depuis les systèmes sources ne sont plus un problème.

Ensuite viennent les technologies de streaming, comme Amazon Kinesis et Apache Spark. La plupart des entreprises transfèrent leurs données par lots massifs ; généralement, cela se produit une fois par jour. Les technologies de streaming permettent aux données d'être ingérées par petits morceaux à très grande échelle. Par exemple, SONOS, le fabricant de haut-parleurs, traite 1 milliard d'événements par semaine à l'aide d'Amazon Kinesis. Il est impensable de devoir attendre que le lot du jour soit transféré pour savoir où en est votre entreprise.

Des insights interactifs où je veux, comme je veux

De nos jours, les utilisateurs en entreprise doivent multiplier les démarches pour comprendre les informations qui leur sont présentées. Peut-être s'agira-t-il de fouiller leur boîte de réception pour retrouver un rapport joint à un e-mail. Ou de se connecter au système de reporting pour télécharger un fichier PDF, puis de se rendre compte qu'ils doivent copier-coller les données dans Excel pour comprendre de quoi il s'agit. Nous devons cesser de contraindre les utilisateurs à effectuer ce genre de parcours du combattant pour obtenir les données et les informations dont ils ont besoin. Le cri de ralliement des utilisateurs devrait être le suivant : « Apportez-nous les données sous la bonne forme, avec les bons outils et au bon moment » !

Des logiciels tels que Tableau, Amazon QuickSight et autres ont fait évoluer la situation en prenant en compte l'expérience des utilisateurs lors de leurs interactions avec les données. Cependant, j'ai constaté que dans la plupart des entreprises, de nombreux outils étaient nécessaires pour répondre aux besoins des utilisateurs. Il peut s'agir d'Amazon QuickSight intégré dans un portail d'intelligence d'affaires, ou simplement d'un classeur Tableau envoyé par e-mail. AWS vous apporte à la fois des outils de stockage de données et d'intelligence d'affaires, le tout via un modèle de paiement à l'utilisation. Cela permet aux entreprises d'essayer de nombreux outils d'intelligence d'affaires sans avoir à trop investir dans les infrastructures et les licences.

Nous devons cesser de contraindre les utilisateurs à effectuer ce genre de parcours du combattant pour obtenir les données et les informations dont ils ont besoin."

Le meilleur algorithme au monde n'a de valeur que s'il peut être intégré aux processus métier."

Dans votre entreprise, les scientifiques des données ne doivent pas être ignorés. Jupyter Notebook a connu un essor fantastique dans la communauté de la science des données ; l'outil s'occupe à la fois de gestion de contenu, d'exécution de code et de visualisation. Il s'agit d'un outil très puissant, autant pour partager des connaissances que pour documenter et exécuter des algorithmes de machine learning. Amazon SageMaker est un environnement de notebook géré qui prend en charge le plus gros du travail pour vous et vos scientifiques des données.

L'intelligence intégrée à l'entreprise

L'intelligence artificielle et le machine learning font fureur de nos jours, et à juste titre. Les progrès réalisés dans les systèmes de machine learning, associés à l'utilisation de serveurs spécialisés utilisant des unités de traitement graphique (GPU), permettent toutes sortes de nouvelles fonctionnalités, comme par exemple la conduite autonome. Bien sûr, pour former des modèles de machine learning, de vastes quantités de données sont nécessaires (d'où les points relatifs aux lacs de données que j'ai abordés plus haut). Les sociétés commencent déjà à tirer parti de ces capacités d'IA/ML pour générer de nouveaux résultats auparavant impossibles, comme l'amélioration des résultats cliniques basés sur l'imagerie rétinienne, ou encore la prédiction des coupures de courant ou des pannes matérielles sur le terrain. Les entreprises peuvent renforcer leurs capacités organisationnelles en matière d'IA/ML en laissant AWS se charger du gros du travail, car cette technologie n'est plus de l'ordre de la science-fiction.

Un dernier point sur lequel j'aimerais insister, c'est que le meilleur algorithme au monde n'a de valeur que s'il peut être intégré aux processus métier. Le plus souvent, créer un modèle de données est la partie la plus facile ; le vrai défi consiste à l'intégrer au moteur de votre police d'assurance ou à votre plate-forme de vente au détail, car ces systèmes ne permettent généralement pas d'intégrer des sources de données ou des API externes. C'est donc l'occasion d'envisager la migration de ces systèmes vers le cloud, afin de tirer parti de tous les services disponibles pour les moderniser ou les réorganiser.

S'organiser pour les insights

La mise en place d'une capacité analytique avancée dans votre entreprise n'est pas affaire que de technologie. Souvent, les plus gros défis que rencontre une entreprise proviennent de l'entreprise elle-même : ses processus, sa gouvernance et ses ressources humaines. Alors, que devez-vous garder à l'esprit pour réussir lorsque vous organisez votre investissement dans l'analytique ?

Commencer par un centre d'excellence analytique

Pour que la stratégie et les bonnes intentions entraînent des progrès significatifs, l'une des premières étapes consiste à identifier et à choisir un leader et une équipe chargés de diriger la transition et à créer un centre d'excellence (COE) analytique. L'équipe commence généralement petit, avec quelques rôles interfonctionnels pour l'amorcer ; elle grandit ensuite peu à peu pour répondre à davantage de besoins.

De nombreuses grandes entreprises ont déjà mis en place des organisations de services partagés qui se chargent de l'intelligence d'affaires ou du reporting. Ces organisations peuvent incorporer dans le centre d'excellence analytique des rôles techniques et commerciaux. Semblables aux infrastructures IT, elles ne doivent pas seulement fournir des talents, mais aussi être un moteur et un promoteur clé de l'effort. Car avec le temps, ces organisations de services partagés devront évoluer pour s'adapter ou s'intégrer au centre d'excellence analytique. Les rôles de départ sont souvent ceux d'ingénieur et d'architecte des données, d'analyste d'intelligence d'affaires et de scientifique des données. Le groupe doit être dirigé par une personne capable de travailler dans plusieurs organisations, succursales et groupes de back-office, comme la fonction financière ou les TI.

Répondre à tous les besoins de vos clients

L'une des principales transitions psychologiques que doivent opérer les entreprises est le passage de l'état d'esprit « vous devez utiliser notre solution de reporting, et ça va vous plaire » à l'état d'esprit « quels sont vos besoins en matière d'analyse, et comment pouvons-nous vous aider » ? Les organisations de services partagés en matière de reporting sont souvent des revendeurs de rapports qui ne sont pas en position de répondre aux questions difficiles posées par les employés, les dirigeants et les clients.

Par conséquent, lors de la création d'un centre d'excellence analytique, il est important d'établir des principes pour le groupe, qui définiront des attentes en matière d'actions et de décisions.

Le centre d'excellence analytique devra répondre aux besoins de deux types de clients :

- **Les consommateurs de données et d'analyses** : Décideurs, scientifiques des données, analystes de veille économique et développeurs. En règle générale, ces clients souhaitent pouvoir accéder rapidement aux informations et aux données, et se soucient de la qualité des outils et des services mis à leur disposition pour traiter et présenter les données.
- **Les producteurs de données** : Propriétaires d'applications, d'infrastructures et d'appareils qui fourniront des données à la plateforme. Ces clients ont besoin de services spécifiques, comme de pouvoir publier facilement leurs données sur la plateforme d'analyse ou définir un contrat de données. Cela inclut le modèle de domaine des données, la fréquence d'actualisation et la définition des stratégies, par exemple une stratégie de sécurité indiquant qui peut accéder à leurs données.

> "Il ne s'agit pas seulement de posséder tous les derniers outils ; il s'agit de permettre à vos clients d'obtenir facilement ce dont ils ont besoin."

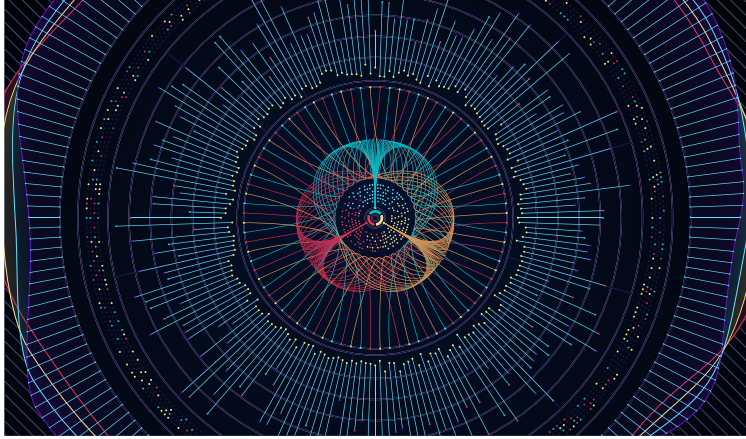
Informations sur l'auteur

Joe Chung est stratège d'entreprise et évangéliste chez Amazon Web Services.

La capacité analytique et la plateforme doivent servir les deux types de clients. Si leurs besoins ne sont pas satisfaits, l'effort analytique n'aura pas de valeur métier. Par conséquent, il est essentiel de disposer d'un mécanisme permettant de cerner les besoins de ces deux types de clients au sein d'un ensemble potentiellement très vaste et diversifié de divisions opérationnelles et de personas. Certaines entreprises mettent en place des conseils consultatifs ou travaillent avec quelques parties prenantes clés afin de mieux répondre aux besoins des clients. Il n'y a pas qu'une seule bonne réponse, mais il est essentiel de disposer d'un mécanisme permettant d'obtenir les commentaires du ou des clients et de hiérarchiser leurs besoins.

Repenser le centre d'excellence

Un centre d'excellence analytique propose un ensemble spécialisé de services cloud axés sur la satisfaction des besoins analytiques. Auparavant, les sociétés de reporting et d'intelligence d'affaires offraient souvent une solution unique pour répondre aux besoins de chacun. À l'ère des technologies en rapide évolution dans les domaines du Big Data, des visualisations enrichies, de la prise de décision automatisée, de l'intelligence artificielle et du machine learning, il n'est tout simplement pas possible de ne disposer que d'une seule pile technologique. Il ne s'agit pas seulement de posséder tous les derniers outils ; il s'agit de permettre à vos clients (producteurs ou consommateurs) d'obtenir facilement ce dont ils ont besoin.



Les centres d'excellence courent le risque de devenir des services de conciergerie. Cela peut convenir pour certains types de requêtes, mais le centre d'excellence peut rapidement se trouver submergé de demandes s'il ne dispose pas de mécanismes évolutifs en libre-service et de processus de hiérarchisation et de gouvernance transparents. Les centres d'excellence analytiques doivent concevoir et mettre en place une plateforme de données en libre-service, sécurisée, exploitable et scalable, dotée d'un écosystème de technologies en constante évolution permettant le traitement, l'analyse et la présentation des informations.

Devenir une entreprise pilotée par les données ne se fait pas du jour au lendemain. Vous devez déjà prendre la bonne direction en identifiant vos problèmes de données, en organisant un plan pour répondre aux besoins de vos clients et en permettant à vos équipes de fournir les bons services aux bons moments.