La *data* dans l'univers bancaire

Par Laurence LE BUZULLIER
Arenium Consulting, associée fondatrice

Les géants du *Web* ont révolutionné l'utilisation et le stockage des données, les *Big Data*. Aujourd'hui, tout est enregistré dans d'énormes *Data Lake*, des conteneurs de données structurées et « non structurées » (telles que les vidéos et les textes), qui viennent en remplacement des anciens *Data Warehouse*. Le *cloud* rend possible l'accès à de telles volumétries de données, tout en réduisant les coûts technologiques supportés par les entreprises.
Mais cette manne financière n'est rien sans intelligence humaine. Pour être valorisée, la donnée doit être rendue intelligible et être exploitée correctement. La donnée a d'autant plus de valeur qu'elle est de qualité, qu'elle circule et qu'elle est mise à jour.
Dans les banques, les données doivent être gouvernées au plus haut niveau, pour que de nouvelles technologies telles que l'Intelligence artificielle puissent être utilisées de façon optimale. La réglementation relative à la donnée se durcit, et les acteurs qui sauront bien l'utiliser en tireront un réel avantage concurrentiel.

« Ce qui n'est pas mesurable n'existe pas », énonce Max Planck, célèbre physicien du XIX^{e} siècle. De tout temps, les chercheurs, mathématiciens et physiciens ont essayé de modéliser le monde sous forme de données et d'équations. Une saison se résume par une température, une pluviométrie et la longueur du jour, une tempête est caractérisée par la vitesse des rafales de vent et sa durée, une étoile se caractérise par sa luminosité, sa masse et sa température…

Les **données** sont des descriptions de la réalité, à travers un prisme. Seules, elles n'ont que peu de valeur. Elles sont indiscutables dans leur individualité, mais difficilement interprétables dans leur unicité, en l'absence d'élément de comparaison. La valeur que peut avoir une donnée naît de sa circulation, de son accumulation, de son échange et des croisements dans lesquels elle intervient.

Aujourd'hui, les données caractérisant le monde sont devenues *Big,* de par le volume des données générées chaque jour. Chaque objet connecté participe davantage à cette mise en statistiques du monde qui nous entoure, à la quantification de celui-ci. L'homme devient une série de chiffres, qui le mesurent et l'analysent. Le nombre de pas que nous faisons quotidiennement, ce que l'on mange, la façon dont on dort, notre rythme cardiaque… tout est enregistré, stocké et analysé. Les géants du *Web*, apparus au milieu des années 2000, ont révolutionné l'utilisation et le stockage des données.

Aujourd'hui, ces données se composent aussi des photos que nous postons, des textes que nous produisons, de notre activité au sens large : tout un ensemble de données « non structurées » qui nous caractérisent. Les données sont des séries de chiffres, bien formatés et structurés, mais aussi des données « non structurées », et donc difficilement interprétables.

Cette manne financière n'est rien sans compétence humaine et sans intelligence. Pour être valorisée, la donnée doit être rendue intelligible.

Les données dans les banques

Dans ce monde numérique en pleine transformation, les banques sont, elles aussi, confrontées à la transformation digitale de leurs activités. Elles sont aujourd'hui dans l'obligation de s'adapter à cette révolution technologique, sous peine de disparaître. Mais pour pleinement réussir leur transformation digitale, les banques doivent accompagner leur mutation d'une approche *Data Centric*. D'une organisation aujourd'hui encore principalement centrée sur le produit, les banques doivent évoluer vers une organisation centrée sur le client, en s'appuyant, en tout premier lieu, sur la donnée. Les données représentent pour les banques une véritable richesse, une richesse souvent mal connue et surtout mal exploitée. Si l'on ne sait pas traiter correctement la donnée, celle-ci perd toute sa valeur et fausse la prise de décision.

En parallèle de cette évolution technologique inévitable, la réglementation bancaire relative à la donnée se renforce constamment. La fiabilité des *reportings,* la protection de la vie privée, la lutte contre le blanchiment et le financement du terrorisme sont autant de sujets qui sont au cœur des préoccupations des autorités.

Ainsi, au lieu de considérer la réglementation comme une contrainte, les banques ont tout à gagner à l'appréhender

comme une véritable opportunité de remettre à plat l'ensemble de leur organisation et de leur gouvernance de la donnée.

Le Comité de Bâle a publié le 9 janvier 2013 un ensemble de principes sous la dénomination BCBS 239, dont l'objectif est d'imposer aux banques une amélioration de leur capacité de production et de fiabilisation des *reportings* réglementaires. L'Europe a par ailleurs publié le RGPD (Règlement général sur la protection des données personnelles) en avril 2016, qui, mis en application en mai 2018, vise à renforcer la protection des données personnelles, obligeant les entreprises à définir une gouvernance de la donnée, et cela dès la conception d'un nouveau produit ou d'un nouveau service. Enfin, les banques sont confrontées à des obligations croissantes sur leurs procédures de connaissance client (*Know Your Customer,* KYC) et de lutte contre le blanchiment et le financement du terrorisme (LCB-FT).

La sécurité autour de la donnée doit également être renforcée : toute fuite de données personnelles des clients doit leur être signalée dans les 72 heures ; l'impact sur la banque est important en risque d'image. Non seulement, une fuite de données porte un véritable préjudice à la banque par la divulgation d'informations confidentielles à des tiers, mais surtout cela entame la confiance que les clients et les autorités peuvent avoir dans l'établissement.

Toutes ces réglementations doivent être considérées par les banques comme un tout, elles ne doivent pas être prises comme des contraintes isolées. Il est intéressant pour les banques d'intégrer le respect de ces réglementations dans un projet global d'entreprise construit autour des données. Ce projet, qui doit dépasser la seule contrainte réglementaire, doit permettre aux banques un meilleur pilotage de leur activité, un gain de temps ainsi qu'une opportunité pour elles d'utiliser de nouvelles technologies prometteuses telles que l'Intelligence artificielle. Ainsi, il apparaît aujourd'hui comme vital pour les banques de mettre en place une vraie stratégie autour de la donnée.

La gouvernance de la donnée dans les banques

Un programme de gouvernance de la donnée correspond à l'organisation de l'ensemble des procédures mises en place au sein de la banque afin d'encadrer la collecte des données et leur utilisation. Ainsi, le programme de gouvernance de la donnée doit s'appliquer non pas à l'échelle d'un service ou d'un *process,* mais à l'ensemble de la banque. Il s'agit pour la banque de respecter les obligations légales et de définir un cadre interne afin d'optimiser l'utilisation des données : la gouvernance de la donnée doit être au service de la prise de décision et non pas s'exercer au nom de la donnée elle-même.

Les banques sont très souvent historiquement organisées en silo : les métiers existent les uns à côté des autres, les fonctions support sont, quant à elles, isolées chacune de leur côté. Cependant, la donnée est, par nature, transversale. Elle circule d'une activité à une autre, en subissant des transformations. Le cycle de vie complet de la donnée doit être maîtrisé, d'un bout à l'autre de la chaîne de production, à travers les différentes organisations. La donnée est saisie par un opérateur en début de chaîne, elle va se transformer et alimenter les *reportings* destinés au *marke-*

ting, au risque ou à la finance en cours de vie, pour terminer à la direction générale, en bout de chaîne, où elle va aider à la prise de décision. Il ne faut pas oublier le grand précepte "*Garbage in, garbage out* !"

La mise en place d'une stratégie d'entreprise doit être effectuée autour de la donnée, et ce au plus haut niveau de la banque.

Du *Data Warehouse* au *Data Lake*

Conçu dans les années 1990 pour les entreprises, le *Data Warehouse* permettait de réconcilier dans une architecture centralisée des informations issues des différents systèmes de gestion de l'entreprise. Il était modélisé au niveau de granularité le plus fin et permettait d'historiser les données sur de longues périodes. Pour accompagner les métiers dans leur exploitation des données, le *Data Warehouse* était accompagné de *Data Marts,* orientés métier, qui permettaient de répondre plus spécifiquement à une problématique fonctionnelle. Les données stockées dans un *Data Mart* étaient en règle générale précalculées et agrégées pour faciliter la restitution prédéfinie ou à la demande, dans des délais raisonnables pour les utilisateurs.

Le *Data Warehouse* ne permettait pas de stocker des données « non structurées », telles que des vidéos, des photos ou des textes bruts.

Ainsi, pour pallier cette difficulté, est apparu en 2014 le concept du *Data Lake*. « Le contenu du *Data Lake* est approvisionné par une source alimentant le lac, et les multiples utilisateurs du lac peuvent venir y plonger pour examiner ou prendre des échantillons », explique James Dixon, le CTO de Pentaho [1], créateur de l'expression *Data Lake*.

Les données du *Data Lake* sont des données brutes, qui sont stockées dans des formats très peu transformés. Chacun doit pouvoir y puiser ce qui l'intéresse, pour l'analyse statistique ou pour le *reporting.*

Mais très vite l'apparition des *Data Lake* a fait émerger un problème, à savoir que, sans gouvernance, un lac de données non structurées n'est accessible qu'à un nombre très restreint d'utilisateurs, aux seuls initiés, à ceux qui savent nager... Les données entreposées sans contrôle, sans qualité certifiée, sans principes de gouvernance sont très difficilement utilisables.

Le modèle d'un *Data Lake* statique, avec des spécialistes accédant aux données pour fournir des *reportings,* est un modèle déjà dépassé. Les données doivent être intégrées directement dans les *process* des banques, pour orienter les clients et la banque en temps réel, générer des alertes, déclencher des actions de maintenance corrective…

Le *cloud*

Tout est rendu possible par le développement du *cloud,* qui s'est opéré en parallèle aux développements du *Big Data* et de l'Intelligence artificielle. Il permet à l'entreprise de développer des applications sécurisées, pilotées par les données, et de rester à la pointe de la technologie. La technologie évolue si rapidement qu'il est compliqué pour les équipes informatiques internes de faire évoluer au même rythme les services existants. Le *cloud* permet d'assurer la sécurité des données et des traitements, et de mutualiser les développements technologiques. Face à l'explosion du *cloud* dans le secteur régulé, l'EBA (European Banking Authority) a publié, fin 2017, son "Final Report on Recommendations on Cloud Outsourcing". Chaque établissement doit ainsi déterminer la matérialité de l'*outsourcing*, pouvoir auditer le *provider* et définir les niveaux de contrôle adéquats, comme pour toute prestation essentielle externalisée.

Par ailleurs, de nouvelles technologies impliquent de nouveaux risques. L'externalisation des données et des traitements des banques vers le *cloud* crée un risque, qui peut être qualifié de systémique (cybercriminalité, indisponibilité…). Le *cloud,* en tant que système informatique, est faillible par définition. L'ANSSI (Agence nationale de la sécurité des systèmes d'information) a mis en place un visa de sécurité afin de qualifier les prestataires de services d'informatique en nuage.

Les plateformes industrielles

La donnée peut être considérée comme une matière première, mais c'est une matière première bien particulière. Elle n'est pas rare (elle est même en perpétuelle expansion), elle reste disponible quand on l'utilise et elle peut être utilisée autant de fois que de besoin, et ce sans perdre de sa valeur (et par autant de systèmes).

La valeur de la donnée augmente avec sa fiabilité, et donc avec le nombre d'occurrences la caractérisant. La donnée a d'autant plus de valeur qu'elle circule, qu'elle est croisée avec d'autres données, qu'elle est mise à jour.

Certains industriels ont bien compris ce principe et ont décidé de mettre en commun leurs données pour gagner ensemble en performance et en fiabilité. Par exemple, Easy Jet a confié sa maintenance prédictive à Skywise, la plateforme de données d'Airbus, qui, en collectant les données de milliers d'avions, veut en améliorer l'exploitation.

Dans le domaine de la banque, l'association GCD (Global Credit Data), créée en 2004, qui regroupe cinquante-deux banques dans le monde (données 2018), œuvre à une mise en commun de leurs données de défauts et de pertes dans le but d'améliorer les performances des modèles internes de risque de crédit de chacun de ses membres.

Les données de la connaissance client ou de la fraude auraient également tout intérêt à être mutualisées par les établissements pour diminuer les coûts de traitement et améliorer la performance des dispositifs de l'ensemble des acteurs du marché. Plusieurs possibilités peuvent

(1) Pentaho se décrit comme « le premier grand fournisseur de solutions décisionnelles à même de proposer des fonctions pour le Big Data *».*

Photo © Romain GAILLARD/REA

Leader mondial du décisionnel et des solutions de *Business Analytics,* SAS France présente les nouvelles fonctionnalités offertes par la solution SAS Visual Analytics en matière de visualisation de données et d'analyse prédictive sur iPad, notamment en ce qui concerne le *Big Data*.

« La *dataviz* englobe toutes les techniques de *Data Visualisation*. Ces techniques permettent de présenter les données de manière visuelle, sous forme de graphiques plus lisibles que de longs tableaux chiffrés. »

être envisagées, à l'instar des « fichiers positifs [2] » de la plupart des pays européens. Les données pourraient être centralisées par un tiers de confiance, telle la Banque de France, ou un organisme privé. On pourrait également envisager la coexistence de plusieurs acteurs, concurrents, mais assurant un minimum d'échanges entre eux.

Les agrégateurs de comptes bancaires peuvent également avoir un rôle à jouer dans la mise en commun d'informations. Ces nouveaux services proposent aux consommateurs de regrouper dans une même application l'ensemble de leurs comptes bancaires détenus dans différents établissements.

La nouvelle directive européenne sur les services de paiements (DSP2), entrée en vigueur en janvier 2018 pour une application définitive en septembre 2019, oblige d'ouvrir l'accès aux informations sur les comptes bancaires *via* un canal de communication sécurisé à des acteurs tiers, tels que les agrégateurs, par exemple (sur demande de leurs clients). Le paysage bancaire est actuellement en pleine mutation et continuera d'évoluer sur ces problématiques dans les années à venir.

Le *Big Data* n'est rien sans intelligence

On parle souvent du *Big Data* en faisant référence aux « 4V », voire « 5V », correspondant aux éléments clés le caractérisant, à savoir Volume, Variété, Vélocité, Véracité (et Valeur). Mais le *Big Data* sans intelligence n'est rien. Les données n'ont de sens que si elles sont mises au service de l'interprétation, de la décision et de l'action. L'« Analytics » peut être résumé comme l'ensemble des techniques permettant de faire parler les données. Ces techniques statistiques ne sont pas nouvelles, mais l'explosion des capacités de traitement des ordinateurs a permis à la statistique inférentielle de se développer de façon exponentielle ces dernières années. Le *credit scoring,* par exemple, a été développé aux États-Unis à la fin des

(2) On appelle « fichier positif » un fichier recensant les crédits détenus par des clients français, dans tous les établissements. Ce fichier est en place dans la plupart des pays européens. En France, seuls les incidents de paiement sont recensés, dans un « fichier négatif ».

années 1960, et est utilisé en France par des spécialistes du crédit à la consommation depuis le milieu des années 1970. Mais, aujourd'hui, les techniques d'Intelligence artificielle permettent non seulement de prévenir les risques de défauts bancaires dès l'octroi des crédits, mais aussi de détecter des mouvements suspects en temps réel, de dialoguer avec les clients, de réaliser des opérations bancaires simples… Ainsi, si les bases de l'Intelligence artificielle existent depuis les années 1960, la puissance informatique a permis de passer, par exemple, d'un « arbre de décision » à une « forêt aléatoire [3] »…

Pour finir, nous pouvons même ajouter un sixième « V », indispensable au *Big Data,* qui est le « V » de Visualisation. La *dataviz* englobe toutes les techniques de *Data Visualisation*. Ces techniques permettent de présenter les données de manière visuelle, sous forme de graphiques plus lisibles que de longs tableaux chiffrés. Une bonne représentation graphique doit mettre immédiatement en évidence le message délivré, permettant soit de pousser l'analyse plus avant, soit d'en tirer une conclusion et, éventuellement, de déclencher une action. Avec l'ère numérique, la multiplicité des informations et l'accélération du temps de la décision, ces techniques de *storytelling* sont devenues essentielles : il faut raconter une histoire, donner du sens aux informations.

Les données sont un outil primordial au service de la banque. Les acteurs qui sauront bien les utiliser en tireront un réel avantage concurrentiel.

(3) L'algorithme des forêts aléatoires effectue un apprentissage sur de multiples arbres de décision (jusqu'à plusieurs centaines) entraînés sur des sous-ensembles de données légèrement différents à chaque fois.