CONCOURS EXTERNE D'ADMINISTRATEUR TERRITORIAL

SESSION 2014

Composition portant sur les technologies de l'information et de la communication

EPREUVE N° 33

Durée : 5 h
Coefficient : 2

SUJET : Valorisation des données publiques volumineuses (Big data)

En avril 2013, le Président de la République a installé la commission Innovation 2030. Celle-ci a été chargée de sélectionner des ambitions fortes, reposant sur des innovations majeures, pour assurer à la France, prospérité et emploi sur le long terme.

Sept ambitions ont ainsi été proposées, et parmi celles-ci, la valorisation des données massives (Big data). Ce domaine concerne autant les entreprises que les administrations et parmi ces dernières, les collectivités locales disposent déjà d'un volume de données considérable. Pourtant, seule une très petite partie de ces données est actuellement exploitée.

Le Président de l'agglomération de GRANDE est également maire de la ville centre depuis 2008. Durant son premier mandat, il a ouvert un certain nombre de données au public (Open Data) et s'intéresse maintenant à l'analyse de données massives (Big data) pour trois raisons :

- **favoriser les nouvelles technologies et les entreprises innovantes ;**
- **valoriser les données publiques afin de mieux cerner les besoins des administrés et des entreprises ;**
- **associer les directions générales et les services autour d'un projet commun.**

Suite à votre réussite au concours d'administrateur territorial, vous êtes recruté(e) par la communauté d'agglomération de GRANDE sur un poste de chargé de mission « agglomération numérique » et directement rattaché(e) au directeur général des services.

En vous appuyant sur les documents joints et vos propres connaissances, il vous est demandé :

- **de rédiger une note d'éclairage sur la valorisation des données massives (Big data) : enjeux, objectifs, contexte et opportunités, contraintes et risques ;**

- **de proposer une feuille de route adaptée à la communauté d'agglomération de GRANDE (objectifs, organisation et moyens, pilotage et conduite du changement.**

DOCUMENTS JOINTS

NOTA :

- 2 points seront retirés au total de la note sur 20 si la copie contient plus de 10 fautes d'orthographe ou de syntaxe.
- **Les candidats ne doivent porter aucun signe distinctif sur les copies** : pas de signature (signature à apposer uniquement dans le coin gommé de la copie à rabattre) ou nom, grade, même fictifs. Seuls la date du concours et le destinataire, (celui-ci est clairement identifié dans l'énoncé du sujet) sont à porter sur la copie.
- Les épreuves sont d'une durée limitée. Aucun brouillon ne sera accepté, la gestion du temps faisant partie intégrante des épreuves.
- Lorsque les renvois et annotations en bas d'une page ou à la fin d'un document ne sont pas joints au sujet, c'est qu'ils ne sont pas indispensables.

Document N°1 Big data - Un extrait de l'article de Wikipédia, l'encyclopédie libre – avril 2014

.....

Les ***big data***, littéralement les **grosses données**[2,1], parfois appelées *données massives*[4], est une expression anglophone utilisée pour désigner des ensembles de données qui deviennent tellement volumineux qu'ils en deviennent difficiles à travailler avec des outils classiques de gestion de base de données ou de gestion de l'information. L'on parle aussi de *datamasse*[5] en français par similitude avec la biomasse.

Dans ces nouveaux ordres de grandeur, la capture, le stockage, la recherche, le partage, l'analyse et la visualisation des données doivent être redéfinis. Les perspectives du traitement des big data sont énormes et pour partie encore insoupçonnées ; on évoque souvent de nouvelles possibilités en termes d'exploration de l'information diffusée par les médias[6], de connaissance et d'évaluation, d'analyse tendancielle et prospective et de gestion des risques (commerciaux, assuranciels, industriels, naturels) et de phénomènes religieux, culturels, politiques[7], mais aussi en termes de génomique ou métagénomique[8], pour la médecine (compréhension du fonctionnement du cerveau, épidémiologie, écoépidémiologie...), la météorologie et l'adaptation aux changements climatiques, la gestion de réseaux énergétiques complexes (via les smartgrids ou un futur « *internet de l'énergie* »...) l'écologie (fonctionnement et dysfonctionnement des réseaux écologiques, des réseaux trophiques avec le GBIF par exemple), ou encore la sécurité et la lutte contre la criminalité[9].

Certains supposent qu'il pourrait aider les entreprises à réduire les risques et faciliter la prise de décision, ou créer la différence grâce à l'analyse prédictive et une « expérience client » plus personnalisée et contextualisée.

Divers experts, grandes institutions (comme le MIT[10] aux États-Unis), administrations[11] et spécialistes sur le terrain des technologies ou des usages[12] considèrent le phénomène *big data* comme l'un des grands défis informatiques de la décennie 2010-2020 et en ont fait une de leurs nouvelles priorités de recherche et développement.

Dimensions des *big data*

Le Big Data s'accompagne du développement d'applications à visée analytique, qui traitent les données pour en tirer du sens. Ces analyses sont appelées Big Analytics [13] ou «broyage de données». Elles portent sur des données quantitatives complexes avec des méthodes de calcul distribué.

En 2001, un rapport de recherche du META Group (devenu Gartner)[14] définit les enjeux inhérents à la croissance des données comme étant tri-dimensionnels : les analyses complexes répondent en effet à la règle dite «des 3V» (volume, vélocité et variété[15]). Ce modèle est encore largement utilisé aujourd'hui pour décrire ce phénomène[16].

Le taux de croissance annuel moyen mondial du marché de la technologie et des services du Big Data sur la période 2011-2016 devrai être de 31.7%. Ce marché devrait ainsi atteindre 23,8 milliards de dollars en 2016 (d'après IDC mars 2013).

Le Big Data devrait également représenter 8% du PIB européen en 2020 (AFDEL février 2013).

Volume

Le volume des données stockées aujourd'hui est en pleine expansion. Selon une étude IDC sponsorisée par EMC Gartner, les données numériques créées dans le monde seraient passées de 1,2 zettaoctets par an en 2010 à 1,8 zettaoctets en 2011, puis 2,8 zettaoctets en 2012 et s'élèveront à 40 zettaoctets en 2020[17]. À titre d'exemple, Twitter génère à l'heure actuelle 7 teraoctets de données chaque jour et Facebook 10 teraoctets[18].

Ce sont pourtant les installations scientifiques qui produisent le plus de données. De nombreux projets, de dimension pharaonique, sont ainsi en cours. Le radiotelescope "Square Kilometre Array" par exemple, produira 50 teraoctets de données analysées par jour, à un rythme de 7 000 teraoctets de donnée brutes par seconde[19]!

Variété

Le volume des Big Data met les data centers devant un réel défi : la variété des données. Il ne s'agit pas de données relationnelles traditionnelles, ces données sont brutes, semi-structurées voire non structurées (cependant, les données non-structurées devront, pour utilisation, être structurées[20]). Ce sont des données complexes provenant du web (Web Mining), au format texte (Text Mining) et images (Image Mining). Elles peuvent être publiques (Open Data, Web des données), géo-démographiques par îlot (adresses I.P), ou relever de la propriété des consommateurs (Profils 360°)[réf. nécessaire]. Ce qui les rend difficilement utilisables avec les outils traditionnels.

La démultiplication des outils de collecte sur les individus et sur les objets permettent d'amasser toujours plus de données [21]. Et les analyses sont d'autant plus complexes qu'elles portent de plus en plus sur les liens entre des données de natures différentes.

Vélocité

La vélocité représente à la fois la fréquence à laquelle les données sont générées, capturées et partagées.
Des flux croissants de données doivent être analysées en temps réel (Data Stream Mining) pour répondre aux besoins des processus chrono-sensibles[22]. Par exemple, les systèmes mis en place par la bourse et les entreprises doivent être capables de traiter ces données avant qu'un nouveau cycle de génération n'ait commencé, avec le risque, que l'Homme perde une grande partie de la maitrise du système quand les principaux opérateurs deviennent des "robots" capables de lancer des ordres d'achat ou de vente à l'échelle des nanosecondes (Trading Haute Fréquence), sans disposer de tous les critères pertinents d'analyse pour le moyen et long-termes.

Différence avec le Business Intelligence

Si la définition du Gartner en 3V est encore largement reprise (voire augmentée de "V" supplémentaires selon l'inspiration des services marketing), la maturation du sujet fait

apparaitre un autre critère plus fondamental de différence d'avec le Business Intelligence et concernant les données et leur utilisation :

- Business Intelligence : utilisation de statistique descriptive, sur des données à forte densité en information afin de mesurer des phénomènes, détecter des tendances… ;
- Big Data : utilisation de statistique inférentielle, sur des données à faible densité en information[23] dont le grand volume permet d'inférer des lois (régressions….) donnant dès lors (avec les limites de l'inférence) au big data des capacités prédictives[24].

Représentation

Modèles

Les bases de données relationnelles classiques ne permettent pas de gérer les volumes de données du Big Data. De nouveaux modèles de représentation permettent de garantir les performances sur les volumétries en jeu. Ces technologies, dites de Business Analytics & Optimization (BAO) permettent de gérer des bases massivement parallèles[25]. Des patrons d'architecture "Big Data Architecture framework (BDAF)"[26] sont proposés par les acteurs de ce marché comme MapReduce développé par Google et utilisé dans le framework Hadoop. Avec ce système les requêtes sont séparées et distribuées à des nœuds parallélisés, puis exécutées en parallèles (map). Les résultats sont ensuite rassemblés et récuperés (reduce). Teradata, Oracle ou EMC (via le rachat de Greenplum) proposent également de telles structures, basées sur des serveurs standards dont les configurations sont optimisées. Ils sont concurrencés par des éditeurs comme SAP et plus récemment Microsoft[27]. Les acteurs du marché s'appuient sur des systèmes à forte scalabilité horizontale et sur des solutions basées sur du NoSQL (MongoDB, Cassandra) plutôt que sur des bases de données relationnelles classiques[28].

Stockage

Pour répondre aux problématiques Big Data l'architecture de stockage des systèmes doit être repensée et les modèles de stockage se multiplient en conséquence.

- le cloud computing : l'accès se fait via le réseau, les services sont accessibles à la demande et en libre service sur des ressources informatiques partagées et configurables[29]. Les services les plus connus sont ceux de Google BigQuery, Big Data on Amazon Web Services, microsoft Windows Azure.

- les super calculateurs hybrides : Les HPC pour High Performance Computing, qu'on retrouve en France dans les centres nationaux de calculs universitaire tels quel'IDRIS, le CINES, mais aussi au CEA ou encore le HPC-LR[30]

Applications des Big Data

Les Big Data trouvent une application dans de nombreux domaines : De grands programmes scientifiques (CERN28 Mastodons), de grandes entreprises (IBM29,Amazon Web Services, BigQuery, SAP HANA) des entreprises spécialisées (Teradata, Jaspersoft30, Pentaho31...) de l'Open Source (Apache Hadoop, Infobright32, Talend33...) et des Start-up (aleph-networks[31],Bionatics[32], Hariba Médical[33], SafetyLine[34], KwypeSoft[35], Vigicolis):

Recherche scientifique

Les expériences du Large Hadron Collider représentent environ 150 millions de capteurs délivrant des données 40 millions de fois par seconde. Il y a autour de 600 millions de collisions par seconde, et après filtrage, il reste 100 collisions d'intérêt par seconde. En conséquence, il y a 25 Po de données à stocker chaque année, et 200 Po après réplication[36,37,38].

Quand le Sloan Digital Sky Survey (SDSS) a commencé à collecter des données astronomiques en 2000, il a amassé plus de données en quelques semaines que toutes les données collectées dans l'histoire de l'astronomie. Il continue à un rythme de 200 Go par nuit, et a aujourd'hui stocké plus de 140 teraoctets d'information. Des prévisions annoncent que le Large Synoptic Survey Telescope, dont la mise en route est prévue en 2015, amassera ce même montant tous les cinq jours[39].

Décoder le génome humain a originellement pris 10 ans, cela peut désormais être fait en moins d'une semaine : les séquenceurs d'ADN ont progressé d'un facteur 10 000 les dix dernières années, soit 100 fois la loi de Moore (100 environ sur 10 ans)[40]. En biologie, les approches massives basées sur une logique d'exploration des données et de recherche d'induction sont légitimes et complémentaires des approches classiques basées sur l'hypothèse initiale formulées [41].

Le NASA Center for Climate Simulation (NCCS) stocke 32 Po de données d'observations et de simulations climatiques[42].

Politique

L'analyse de Big Data a joué un rôle important dans la campagne de ré-élection de Barack Obama, notamment pour analyser les opinions politiques de la population[43].

Depuis l'année 2012, le Département de la défense américain investit annuellement sur les projets de Big Data plus de 250 millions de dollars[44].

Le gouvernement américain possède six des dix plus puissants supercalculateurs de la planète[45].

La National Security Agency est actuellement en train de construire le Utah Data Center. Une fois terminé, ce data center pourra supporter des yottaoctets d'information collectés par la NSA sur internet.

En 2013, le Big Data faisait partie des 7 ambitions stratégiques de la France déterminées par la Commission innovation 2030[46]

Secteur privé

Walmart traite plus d'un million de transactions client par heure, celles-ci sont importées dans des bases de données dont on estime qu'elles contiennent plus de 2,5 Po d'information[39].

Facebook traite 50 milliards de photos.

D'une manière générale le data mining de Big Data permet l'élaboration de profils clients dont on ne supposait pas l'existence[47].

Perspectives et évolutions

L'un des principaux enjeux de productivité du Big Data dans son évolution va porter sur la logistique de l'information, c'est à dire sur comment garantir que l'information pertinente arrive au bon endroit au bon moment. Il s'agit d'une approche micro-économique. Son efficacité dépendra ainsi de celle de la combinaison entre les approches micro- et macro-économique d'un problème.

Selon une étude IDC, les données numériques créées dans le monde atteindraient 40 zettaoctets d'ici 2020[48]. A titre de comparaison, Facebook générait environ 10 teraoctets de données par jour au début 2013. Le développement de l'hébergement massif de données semble avoir été accéléré par plusieurs phénomènes simultanément: la pénuerie de disques durs due aux inondations en Thaïlande en 2011, l'explosion du marché des supports mobiles (smartphones et tablettes notamment), etc. Ajouté à cela, la démocratisation du cloud-computing de plus en plus proche, grâce à des outils comme Dropbox, amène le big data au centre de la logistique de l'information.

Afin de pouvoir exploiter au maximum le Big Data, de nombreuses avancées doivent être faites, et ce en suivant trois axes :

Modélisation de données

Les méthodes actuelles de modélisation de données ainsi que les systèmes de gestion de base de données ont été conçus pour une utilisation à des fins commerciales de l'information. La fouille de données a des caractéristiques fondamentalement différentes et les technologies actuelles ne permettent pas de les exploiter. Dans le futur il faudra des modélisations de données et des langages de requêtes permettant :

- une représentation des données en accord avec les besoins de plusieurs disciplines scientifiques ;
- de décrire des aspects spécifiques à une discipline (modèles de métadonnées) ;
- de représenter la provenance des données ;
- de représenter des informations contextuelles sur la donnée ;
- de représenter et supporter l'incertitude ;
- de représenter la qualité de la donnée[49].

Gestion de données

Le besoin de gérer des données extrêmement volumineuses est flagrant et les technologies d'aujourd'hui ne permettent pas de le faire. Il faut repenser des concepts de base de la gestion de données qui ont été déterminés dans le passé. Pour la recherche scientifique, par exemple, il sera indispensable de reconsidérer le principe qui veut qu'une requête sur un SGBD fournisse une réponse complète et correcte sans tenir compte du temps ou des ressources nécessaires. En effet la dimension exploratoire de la fouille de données fait que les scientifiques ne savent pas nécessairement ce qu'ils cherchent. Il serait judicieux que le

SGBD puisse donner des réponses rapides et peu coûteuses qui ne seraient qu'une approximation, mais qui permettraient de guider le scientifique dans sa recherche[49].

Dans le domaine des données clients, il existe également de réels besoins d'exploitation de ces données, en raison notamment de la forte augmentation de leur volume des dernières années[50]. Le big data et les technologies associées permettent de répondre à différents enjeux tels que l'accélération des temps d'analyse des données clients , la capacité à analyser l'ensemble des données clients et non seulement un échantillon de celles-ci ou la récupération et la centralisation de nouvelles sources de données clients à analyser afin d'identifier des sources de valeur pour l'entreprise.

Outils de gestion des données

Les outils utilisés à l'heure actuelle ne sont pas en adéquation avec les volumes de données engendrés dans l'exploration de Big Data. Il est nécessaire de concevoir des instruments permettant de mieux visualiser, analyser, et cataloguer les ensembles de données afin de permettre une optique de recherche guidée par la donnée[49]. La recherche en Big Data ne fait que commencer. La quantité de data évolue beaucoup plus rapidement que nos connaissances sur ce domaine. Le site *the Gov Lab* prévoit qu'il n y aura pas suffisamment de scientifiques du Data. En 2018, les États-Unis auraient besoin de 140 000 à 190 000 scientifiques spécialisés en Big Data[44].

Document N°2 Extrait du rapport de la commission innovation 2030 présidée par Anne Lauvergeon – juillet 2013

SYNTHÈSE

La Commission Innovation, composée de 20 personnalités aux profils variés, a été installée par le Président de la République le 18 avril 2013.

Il lui a été demandé, par lettre de mission du Premier ministre, de sélectionner, en nombre limité, des ambitions fortes, reposant sur des innovations majeures, pour assurer à la France prospérité et emploi sur le long terme. Son objectif est de stimuler l'innovation au sein des entreprises de toute taille autour de priorités durables. Pour ce faire, la Commission est convaincue qu'il faut éviter la dispersion et le « zapping » pour réussir.

Si l'innovation peut être favorisée par une action volontariste des pouvoirs publics dans la durée à l'exemple d'Airbus, elle naît aussi d'initiatives individuelles et répond à des demandes sociétales. Le rôle de l'État et des collectivités territoriales est alors d'assurer l'existence d'un environnement favorable.

La Commission s'est d'abord concentrée sur le contexte dans lequel la France évolue avant de définir une méthode de travail permettant de présenter des choix stratégiques d'innovation. Pour ce faire, elle a auditionné des personnes de tous horizons, a reçu de très nombreuses contributions et a mené une étude des politiques développées dans différents pays à partir de sources ouvertes et du travail des ambassades.

La définition de ces choix s'appuie sur les attentes sociétales fortes, en croissance – préoccupation pour la planète, vision plus « individualiste » du citoyen-consommateur, responsabilité individuelle accrue, etc. – mais également la prise en compte d'un contexte international complexe - potentiel économique des pays émergents, allongement de la durée de la vie, urbanisation croissante, tensions probables pour l'accès à l'eau potable, à l'énergie et aux matières première, effets croissants du changement climatique. La grille de lecture du monde évolue : le progrès se conjugue avec les notions d'utilité, de sobriété et de bien d'usage. Le besoin de sécurité s'accroît qu'il s'agisse des personnes, des biens ou des informations, en parallèle à une volonté de santé et de bien-être à tout âge. Les innovations de demain devront répondre à ces attentes montantes de la société et arriver au bon moment. A défaut, elles ne rencontreront pas leur marché et resteront sur étagères.

Pour passer du possible au réel, la France dispose de solides atouts même si la concurrence mondiale s'accroît. Le tour d'horizon international de la Commission montre que beaucoup d'États mettent en place des stratégies d'investissement ciblées pour acquérir des positions de leaders dans certains domaines. Il ne suffit plus de disposer de forces dans un domaine. Il faut être à la pointe de l'innovation, présenter des atouts d'excellence, convaincre de la qualité au bon moment et attirer les meilleurs talents dans un contexte de concurrence internationale.

Mais la France présente aussi des handicaps, avec un écosystème culturel et une organisation qui n'incitent pas à l'innovation et sur lequel il faut agir. Fiscalité, contraintes réglementaires, conjoncture morose ou frilosité tout simplement ne facilitent pas la vie des innovateurs. Ce constat n'est pas nouveau. La France a peur d'oser et de prendre des risques. Elle est actuellement l'antépénultième pays en termes de production économique industrielle en Europe.

.....

7 ambitions fortes ont été définies et parmi elles la valorisation des données massives (Big data) proposée page suivante

.....

(...)

Ambition n°7 : La valorisation de données massives (Big Data)

-

La multiplication des données créées par les particuliers, les entreprises et les pouvoirs publics sera porteuse de nouveaux usages et de gains de productivité. La mise à disposition par l'État et par ses opérateurs des données publiques constituera une opportunité pour favoriser l'essor de nouvelles start-up. Ici encore, la France présente de nombreux atouts. L'école française de mathématiques et de statistiques est une des meilleures au monde. Plusieurs entreprises sont leaders de sous-segments.

Ces Ambitions ont été choisies par la Commission sur la base de plusieurs critères, en premier lieu leur capacité à générer de la croissance, des emplois et des exportations. Elles sont à la confluence de marchés majeurs portés par des besoins sociétaux certains et de compétences distinctives françaises. Elles nécessitent des innovations de rupture et constituent pour la Commission un enjeu de souveraineté pour que la France soit durablement une puissance économique prospère. Enfin, elles prennent en compte des évolutions technologiques massives comme la révolution numérique ou l'impact des nouveaux matériaux avec des propriétés avancées.

Dans ce contexte, l'exercice de la Commission s'inscrit en complémentarité du projet de « Nouvelle France industrielle » qui met en oeuvre 34 plans définissant des relais de croissance des filières industrielles sur les marchés d'aujourd'hui. La Commission veut, quant à elle, susciter, d'ici dix ans, des leaders industriels français à l'échelle internationale, dans des secteurs précis, en concentrant les moyens sur des axes clefs.

Ces Ambitions pourront avantageusement s'appuyer sur des consortia européens. Elles s'inscrivent en effet pleinement dans les différents défis sociétaux identifiés par la Commission européenne.

Après plus de deux décennies de gains de productivité très importants dans l'entreprise, les technologies de l'information ont, depuis le début des années 2000, essentiellement bénéficié au grand public, avec l'adoption en masse de l'Internet, des réseaux sociaux ou encore du e-commerce. Ces nouveaux usages ont donné lieu à la naissance des géants comme Google, Yahoo, Facebook ou Amazon, pour ne citer que les plus grands, et a conduit ces derniers à recueillir des quantités de plus en plus considérables de données (moteurs de recherche, ciblage publicitaire, données d'usage, etc.). Les technologies existantes, comme les bases de données relationnelles, se révélant incapables de gérer de telles quantités de données, ces sociétés ont été amenées à développer leurs propres technologies de stockage et de traitement de ces données. Il s'agit là du Big Data.

D'autre part, de nouveaux usages sont apparus en lien avec le développement des applications sur smartphones, notamment dans les transports et la mobilité. Cette évolution met en évidence l'urgence, pour le secteur numérique, de mettre à disposition des développeurs de données d'intérêt général comme les statistiques en tous genres détenues par les pouvoirs publics. C'est ce que l'on appelle les données ouvertes ou « Open Data ». D'autres types de données, détenues par des acteurs privés ou parapublics, sont aussi essentielles au développement des nouveaux usages comme les données de consommation des compteurs électriques ou les informations sur l'état des parkings de vélos dans des systèmes de type Vélib. Il ne s'agit néanmoins pas de données ouvertes.

L'exploitation de ces données massives dont disposent les entreprises et les pouvoirs publics sont porteuses d'applications nouvelles et de gains de compétitivité considérables dans des domaines aussi variés que la santé (gestion des systèmes d'assurance maladie, génomique, épidémiologie, etc.), l'environnement, l'agriculture, le secteur de la banque/assurance, la culture, le tourisme, la publicité en ligne, le marketing, la recherche, l'éducation, les études économiques ou démographiques, la relation client... Des projets émergents comme les « smart cities » ou les « smart grids » génèreront beaucoup d'informations qu'il faudra traiter en temps réel.

La capacité pour les entreprises, les individus et les objets intelligents (robots, interfaces hommes-machines, objets intelligents connectés, capteurs, ...) à exploiter de façon pertinente ces énormes quantités d'informations est un enjeu d'autant plus important que des données issues de secteurs éloignés d'une entreprise peuvent être d'un intérêt primordial pour elle (par exemple, la détection de la propagation d'une épidémie en temps réel par les requêtes sur les moteurs de recherche). Ces nouvelles méthodes de traitement des données permettront également d'accroître l'automatisation, d'agir plus rapidement mais aussi de mieux connaître ses clients.

Cette exploitation des données est donc un enjeu économique indéniable des prochaines années.Mc Kinsey estime qu'en 2025, les Big Data représenteront 5 000 milliards de dollars par an. Les applications seront multiples et concerneront tous les domaines industriels. La valeur ajoutée française de cette filière est estimée à 4,8 Mds € en 2010 avec une croissance d'environ 7% par an, mais avec un impact bien supérieur sur tous les secteurs économiques, et notamment par la « marchandisation » progressive de bases de données (Massive Open Online Courses, par exemple).

Cette évolution technologique rencontre des tendances sociétales de fond. Le citoyen consommateur souhaite avoir accès à de plus en plus d'informations pour décider par lui-même. Il demande également de plus en plus une information personnalisée, c'est-à-dire adaptée à son cas précis. L'information extraite doit donc être individualisée pour répondre à un besoin précis : traiter ses maladies en fonction de son génome et de ses habitudes de vie, apprendre selon son profil et ses ambitions, définir son profil de risque personnel, etc. De telles offres doivent également scrupuleusement respecter la vie privée des individus. L'enjeu est donc non seulement technologique mais aussi législatif et réglementaire pour concilier compétitivité et capacité d'innovation des entreprises avec le respect de la vie privée.

En dehors des enjeux de compétitivité des entreprises, déjà cités, ce secteur comporte aussi des enjeux de souveraineté sur les données, de sécurité nationale (cyber sécurité) et d'exploitation de ce potentiel dans l'administration. Il s'agit aussi de permettre un accès efficace des petites entreprises à l'international, pour en assurer un développement aussi rapide qu'aux États-Unis, et développer une offre française à l'échelle mondiale.

Face aux enjeux économiques que représente la valorisation des données massives, la Commission est persuadée qu'il s'agit d'un enjeu qu'il importe que la France maîtrise d'ici 2025. Différentes temporalités existent. Une partie des évolutions seront incrémentales et ne sont pas envisagées ici. Des ruptures avec des efforts de R&D à une échelle de temps plus long peuvent parallèlement être envisagées.

Pour ce faire, la France peut compter sur plusieurs points forts.

Le système éducatif français forme des ingénieurs généralistes ayant une très bonne maîtrise des mathématiques et des statistiques, nécessaires aux algorithmes capables de traiter des informations hétérogènes et gigantesques. L'école française de mathématiques et de statistiques est ainsi internationalement reconnue comme une des meilleures au monde et nos étudiants sont très recherchés. La recherche publique française présente également un haut niveau d'excellence en la matière.

La France abrite plusieurs sociétés de niveau international, notamment dans le domaine de l'Internet des objets (Withings, Sigfox, Parrot, ...) qui n'ont rien à envier à leurs concurrents, ou encore des sociétés comme Criteo dans le domaine du ciblage publicitaire, qui est l'un des champions mondial, avec une taille déjà très significative. Plusieurs grands groupes sont leaders de sous-segments (Dassault Systèmes, Gemalto, Ingenico, Morphosytèmes,...). Un écosystème dynamique de start-up existe ainsi en France autour de ce sujet. Des pôles de compétitivité du domaine des TIC, qui favorisent les coopérations publiques privées, tels que Systematic, Cap Digital, Images & réseaux ou Solutions communicantes sécurisées, sont un outil de concentration de cet écosystème.

Le statut de jeune entreprise innovante est particulièrement pertinent pour ce domaine.

La France a une tradition de pionnier, avec la Commission nationale de l'informatique et deslibertés (CNIL), dans la gestion raisonnée des données personnelles et, moyennant une réglementation équilibrée, notre pays pourrait devenir le terreau d'innovations d'usage dans le domaine du Big Data.

Enfin, un nombre important de données sont disponibles à l'échelle nationale et ne demandent qu'à être valorisées, l'État français étant construit autour d'une organisation centralisée.

La valorisation des données massives en France fait néanmoins face à plusieurs difficultés.

Il importe tout d'abord d'inventer des solutions innovantes (bases de données en mémoire, nouvelles architectures de traitement, analyse en temps réel, méthodes d'apprentissage automatique, nouveaux modèles de modélisation de données, etc.) et des modèles économiques autour de ces données. La question de l'accès au financement pour la croissance des entreprises du secteur est ainsi fondamentale.

Ensuite, face à ces données, les débats sont nombreux. Il importe d'assurer à la fois la sécurité de ces données et leur accessibilité, la protection de la vie privée et la liberté d'usage. Ainsi, le traitement et l'exploitation des informations numériques ne doivent-ils pas porter atteinte au respect de la vie privée et aux libertés individuelles. En dehors des fichiers qui comportent des données personnelles et qui, en France, sont contrôlés et régulés par la CNIL, toute personne laisse des traces numériques qui peuvent permettre de recueillir des informations sur elle : recherches sur Internet, commandes en ligne, etc.

Des affaires récentes, comme le système de cyber-surveillance PRISM de la NSA américaine, sont révélatrices de la frontière fragile qui existe entre le respect de la vie privée et la nécessité de disposer de technologies avancées (cybersécurité) pour lutter contre le terrorisme, la pédophilie, etc. La personnalisation de l'offre de services, comme l'apparition de bannières publicitaires ciblées sur Internet, présente une valeur ajoutée à la fois pour l'usager d'Internet et pour le vendeur, mais repose sur la collecte d'informations sur les pages consultées par l'utilisateur. L'agrégation et/ou l'anonymisation des données est cruciale. Dans la plupart des cas, il n'est pas nécessaire d'obtenir des informations nominatives. Au-delà, la question de la propriété des données doit être posée.

Si des règles, acceptées à l'échelle internationale, apparaissent clairement nécessaires, pour proscrire la surveillance d'individus en dehors de tout cadre légal, il ne faut pas que celles-ci deviennent une interdiction *a priori* de technologies par la France qui empêcherait les entreprises françaises d'expérimenter et de promouvoir de nouveaux usages.

Propositions de leviers d'actions

1. Ouvrir les données publiques, rendues anonymes, pour favoriser la création de start-up et créer des écosystèmes en France par la valorisation de certains usages à des fins commerciales.
Cette mesure, déjà adoptée notamment en Grande-Bretagne sous le terme d'« Open Data », est gratuite pour l'État et peut permettre une meilleure connaissance des marchés par les entreprises. Tous les secteurs et toutes les infrastructures sont concernés : santé, énergie, transport....

2. Faire changer d'échelle les entreprises françaises en lançant des défis de valorisation de stocks de données massives.

La France, par sa tradition centralisée, dispose de stocks de données de dimension très importante (INSEE, données administratives, sécurité sociale, etc.). Il s'agit de lancer des programmes de valorisation par licence de cinq « stocks » de données massives dont l'analyse pourra apporter une plus-value à l'ensemble de notre société : Pôle emploi, la Sécurité sociale, l'éducation nationale et enseignement supérieur ainsi que les aides à la valorisation du patrimoinetouristique. D'autres défis de

valorisation comme la gestion intelligente de l'énergie peuvent également être envisagés par les pouvoirs publics en lien avec le monde économique.

Par leur masse, l'exploitation de ces données représente un objectif essentiel pour les entreprises participantes et constitue une référence de valeur. Par ailleurs, cette valorisation des données publiques permettra de renforcer l'efficacité de l'action publique par l'exploitation « intelligente » des données considérables dont dispose l'administration et la découverte de nouvelles possibilités d'analyse.

Ponctuellement, l'intervention de l'État pourra aussi se concrétiser par le soutien au développement des start-up du domaine, souvent très consommateur de capital dans les premières phases.

Il importe également de favoriser les start-up qui créent et accumulent des données en propre. Ces entreprises auront en effet un avantage compétitif décisif sur le marché et capteront une part essentielle de la valeur.

3. Créer un droit à l'expérimentation.

L'approche traditionnelle (réglementation et administration de contrôle) est mal adaptée aux constantes du temps des usages qui se développent grâce à ces technologies. Un droit à l'expérimentation doit être reconnu, et encadré par un « observatoire des données ».
Il importe en effet de ne pas légiférer sur ce thème de manière générique. L'usage des données est sectoriel et demande une approche au cas par cas. Cette méthode pourrait être progressivement élargie à l'échelle européenne de manière, dans la mesure du possible, à construire une réglementation commune au niveau européen.

La Commission pense possible, par une approche sectorielle et par type d'usage, de définir une législation et une réglementation pertinente. Il importera de prendre le temps d'observer le développement des nouveaux usages avant de légiférer. L'exemple de la relation de confiance entre les banques et les usagers prouve qu'il est possible d'avoir une approche gagnant-gagnant dans le domaine de la gestion des données personnelles, mais certains systèmes comme le profilage des utilisateurs pour la publicité devront sans doute être gérés de manière spécifique.
De même, il est indispensable d'imposer une étude d'impact économique avant toute législation sur ce sujet, afin de préserver l'équilibre souhaitable entre innovation, compétitivité et respect de la vie privée.

4. Créer un centre de ressources technologiques.

Un « centre de ressources technologiques » dédié pourrait contribuer à abaisser considérablement la barrière à l'entrée que constitue la maîtrise des technologies très complexes du Big Data, et ainsi réduire le time-to-market des « jeunes pousses », maximalisant leurs chances de devenir des leaders mondiaux.

Il s'agirait de mettre à la disposition des acteurs innovants des outils logiciels, des méthodes statistiques ou mathématiques, des jeux de données massives ou des infrastructures de calcul massivement distribuées, permettant de mettre au point très rapidement de nouveaux usages fondés sur les technologies du Big Data. Ce centre de ressources technologiques serait ouvert à tous (start-up comme grand groupes) et chacun pourrait y contribuer.

15

Document N°3 Présentation de la communauté d'agglomération de GRANDE et de sa ville centre

INTRODUCTION

Nombre d'entreprises sont implantées sur le territoire de GRANDE, en plus d'une grande université et d'un centre hospitalier régional universitaire. Depuis plusieurs années, un partenariat étroit s'est développé entre l'université, les entreprises et les collectivités locales.

La Communauté d'agglomération de GRANDE exerce ses compétences sur un territoire urbain d'environ 400.000 habitants. Quinze communes se partagent le territoire communautaire. Les compétences de l'EPCI ont été progressivement étendues. Elles concernent la distribution de l'eau, l'assainissement, la voirie, la circulation, les transports en commun, l'environnement et la gestion des déchets, les piscines, el conservatoire, trois musées, le centre des congrès, ainsi que plusieurs directions ressources (Finances, Ressources Humaines, Administration Générale, Patrimoine, Direction Juridique et Direction des Systèmes d'Information.

La ville centre gère nombre d'activités culturelles (Opéra, Théâtre, musées, Orchestre, Bibliothèques, Archives et des manifestations), l'action sociale, l'équipement des écoles, le développement d'activités sportives et de la jeunesse, les formalités administratives. A ces services, il faut également ajouter les services techniques, la police municipale. L'ensemble des services ont été regroupés sous l'égide d'une structure publique unique.

Ainsi cette administration commune comprend 7 grandes directions (développement urbain, culture, Ecoles-sports-jeunesse-loisirs-Etat-civil, services urbains (transports-voirie-gestion de l'eau,..), grands projets-économie et université, activités sociales, ressources). A chaque grande direction, sont rattachés plusieurs services opérationnels ; la Communauté d'agglomération de GRANDE en compte plus de quarante.

La Direction des Systèmes d'Information a modernisé le fonctionnement des services tant pour les applications classiques de gestion que pour les services à la population. On lui reconnait une mise œuvre efficace du système d'information géographique. Les équipements mobiles ont largement été développés (600 tablettes et smartphones). Plusieurs applications ou systèmes innovants ont été mis en œuvre.

1 – Ouverture des données au public (Open Data)

Depuis deux ans, la Communauté d'agglomération de GRANDE a ouvert au public des jeux de données. On en compte 145 jeux de données qui touchent différents domaines, depuis les noms de voie au transport, des données géographiques et cadastrales, des données financières, des données population (non nominatives).

L'ouverture des données au public n'est pas directement liée à la problématique de valorisation des données volumineuses (Big Data), toutefois, l'Open Data nécessite la consolidation de données multiples et leur nécessaire mise à jour régulière. C'est de plus, très souvent des données de référence.

Le Président de l'EPCI a déjà développé, en lien avec l'université et des petites entreprises des applications innovantes et il a compris qu'il suffit souvent d'un élément déclencheur pour lancer un projet, voire une entreprise. Or, il ressent que dans le domaine des technologies de l'information, il y a beaucoup à faire.

2 – Gestion de la relation citoyen

Ce système, en service depuis plusieurs années, est constitué :

- d'un centre d'appels disposant d'une équipe de 6 agents. Ce centre enregistre les demandes en direct de 7h30 à 19h mais dispose également d'une messagerie ouverte en permanence.
- Toutes les demandes sont instruites et transférées au service compétent
- Le service prépare le travail, intervient et prépare la réponse par courriel ou courrier suivant la demande initiale.
- Une enquête de satisfaction est lancée en même temps que la réponse.

Ce service reçoit environ 40.000 demandes chaque année avec un bon degré de satisfaction ; toutefois aucune analyse de données n'est faite sérieusement

3 – Sécurité

Un système de vidéo-tranquillité est déployé sur l'ensemble de la communauté d'agglomération et un pc sécurité a en charge la supervision. Les données ne sont conservées que sur une durée réglementaire (1 mois) mais le volume quotidien enregistré est impressionnant.

4 – Transports

Tous les trajets des transports publics sont enregistrés, de même que les horaires d'arrivée et de départs des véhicules aux arrêts. Le volume de données résultant est conséquent mais peu exploité et surtout, il n'est pas associé à d'autres critères (périodes de l'année, température, événements, travaux, accidentologie, développement des pistes cyclables, du vélo-cité).

5 – Ville intelligente

Ce concept de développement a été utilisé à plusieurs reprises, sur la gestion de la circulation, sur un site de mobilité (application mobile utilisable sur les smartphones) permettant de se repérer, de trouver le chemin le plus rapide, les horaires de transport, les stations de taxis. Les usagers ont également la possibilité de signaler en temps réel tout incident sur la voie publique.

6 – Gestion des déchets

L'usage des déchetteries est réglementé par carte d'accès (gratuite pour les administrés de l'agglomération) et tous les accès sont tracés et enregistré.

7 – Sites Web

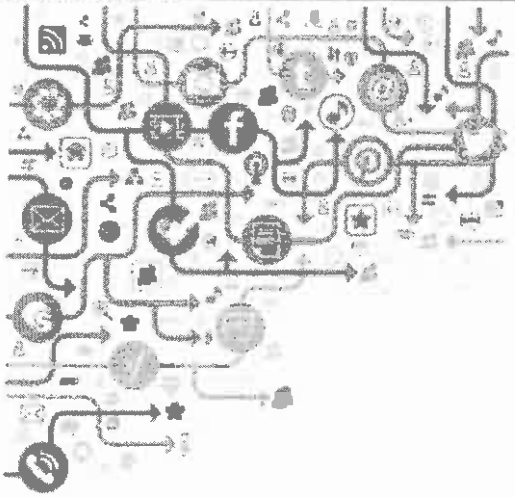
On compte une quinzaine de sites internet qui sont évalués et suivi par la direction de la communication (pages les plus consultées, observations et demandes).

La DSI gère près de 200 applications métiers et dispose d'un volume de données considérable (plus de 80 téraoctets) dans tous les domaines d'activité.

Document N°4 Big data, l'heure est à la valorisation des données

LeMonde Informatique Article de Serge Leblal – avril 2013

Le 05 Avril 2013



Sur le salon Big Data 2013, l'ambiance générale était particulièrement studieuse au coeur du petit écosystème français. Un rendez-vous stimulant pour mieux comprendre comment valoriser ses informations.

Pour sa seconde édition, le salon Big Data, qui a migré de la Cité universitaire de Paris au CNIT, a accueilli un peu plus de 2 500 participants contre 700 l'an dernier. Blandine Laffargue, l'organisatrice de l'évènement, nous a indiqué sa satisfaction quant à la hausse des exposants (50 contre 25 en 2012), et des conférences (100 au lieu de 70). « Nous avons amplifié l'évènement dans tous les domaines », précise B. Laffargue, « avec notamment les ateliers produits où fournisseurs et partenaires pouvaient exposer leurs solutions et confronter leurs idées avec les clients potentiels » pendant 30 minutes. « En un an le marché du big data a vraiment décollé. L'année dernière, il était très difficile de trouver des projets innovants dans ce domaine, alors que cette année nous avons remonté 43 projets. » Mais c'est une fois de plus le Crédit Mutuel Arkea qui a remporté le premier prix des Trophées de l'Innovation 2013 de ce salon. L'année dernière, nous avions déjà remis à Mathias Herberts, « ingénieur disruptif » du Crédit Mutuel Arkea, le **Trophée de l'Innovateur CIO/LMI 2012 pour la mise en oeuvre d'une solution big data transverse en technologies Hadoop**. Le second prix a été attribué à Monster pour son programme Seemore et le troisième à SFR pour un projet géomarketing dynamique.

Dans les allées du salon, nous avons pu discuter avec un grand nombre de start-ups dédiées au big data comme Affini-TECH fondée par Vincent Heuschling, mais aussi des

SSII et des éditeurs. La première bonne surprise sur le salon était toutefois la présence de MapR Technologies, une des trois distributions majeures Hadoop avec celles de Cloudera et de Hortonworks. La seconde était l'ouverture d'une filiale française de MapR avec aux manettes Xavier Guérin, auparavant chez Isilon Systems et Quantum, comme vice-président en charge de l'Europe du Sud et du Benelux, et Aurélien Goujet, auparavant chez Isilon Systems, comme directeur technique Europe du Sud et du Benelux. Jusqu'à présent les trois principaux protagonistes oeuvrant sur le framework Hadoop n'étaient représentés que par leurs partenaires, VirtualScale pour Cloudera, par exemple, avec Sofiane Ammar et Maurice Abecassis.

Ted Dunning de MapR en évangéliste Hadoop

Sur le salon, MapR était très bien représenté grâce à la présence de Ted Dunning, chief application architect, qui a assuré une des conférences du salon. Ce dernier nous a souligné le travail de l'éditeur sur le framework Hadoop avec notamment l'utilisation du système de fichiers NFS (Network File System), associé à un connecteur HDFS pour garder la compatibilité avec le framework d'origine. MapR propose en fait deux versions de sa distribution, une de base dite M3, qui prend en charge le Network File System (NFS) pour assurer un déploiement plus facile avec les systèmes de stockage et de meilleures performances en débit (jusqu'à 20 Gbit/s), et une version dite M5, configurée pour la haute disponibilité (HA). Elle supporte également le multitenancy, ce qui lui confère un certain nombre d'avantages. Le logiciel de gestion peut maintenant supporter les clusters multiples, si bien que les administrateurs peuvent partitionner logiquement un cluster physique et lui attribuer des tâches différentes. Enfin des fonctions de snapshot et de mirrorring sont également de la partie. Ted Dunning s'est félicité de la croissance du big data aux États-Unis , avec une adoption qui explose depuis en un an et des projets qui fleurissent un peu partout dans la finance, la distribution, l'industrie... Le marché initial qui concernait essentiellement les opérateurs web est aujourd'hui totalement transformé.

Après un retard à l'allumage, le marché français du big data commence à se développer. Arnaud Laroche, associé chez Bluestone, nous a indiqué quelques usages très intéressants chez Air France pour la fixation dynamique du prix des billets d'avion et à la Caisse des dépôts pour la valorisation des brevets. Bluestone, qui emploie aujourd'hui 120 personnes, ne craint pas la pénurie de compétences. « En France, le profil des data scientists est différent de celui des États-Unis. Nous avions moins d'ingénieurs en informatique, mais plus de scientifiques et de mathématiciens attirés par ces nouveaux

métiers. » Et la révolution n'est pas que dans les profils, elle est aussi dans les usages. « Aujourd'hui la data devient opérationnelle pour le développement de produits ou de services, notamment des alertes pour la maintenance avec, par exemple, la détection de signaux faibles ».

HP pousse bien sûr Autonomy IDOL

Mais le big data ne se limite pas à Hadoop, des éditeurs poussent leurs propres solutions pour traiter et analyser de grandes quantités de données. Progress Software, par exemple, mettait en avant son travail réalisé chez Turkcell, le 3e opérateur turc avec 20 millions de clients, pour réduire le taux d'attrition avec son moteur CEP. Ce dernier rassemble et traite en temps réel des couches de données issus de plusieurs sources (mobiles, flux sociaux...) pour filtrer et corréler les informations. Un outil de realtime marketing devenu indispensable pour dépasser la simple segmentation marketing.

Enfin, HP était également sur le salon pour mettre en avant ses plates-formes Autonomy IDOL (Intelligent Data Operating Layer) et Vertica. Jean Paul Alibert, directeur général chez HP France en charge de l'innovation, du big data et de la sécurité. « Des trois offres en croissance sur le marché (cloud, sécurité et big data), le big data offre aujourd'hui les plus larges opportunités. Avec Autonomy, nous possédons un outil capable de traiter et marquer des données structurés et non structurées, mais aussi des rich médias avec notamment la reconnaissance de visages et de logos. L'audio peut en outre être retranscrit en texte pour être analysé en temps réel ». Autonomy assure également des fonctions d'analyse de sentiments grâce à la détection de mots clefs dans une conversation et à l'analyse du spectre vocal pour détecter des tensions entre un client et un opérateur dans un centre d'appels. En cas de problème, le client peut être automatiquement basculé sur un manager pour régler le souci. Les principaux POC big data emmenés par HP aujourd'hui concernent la banque et assurance pour mieux cibler les clients via leur relevé bancaire. Et ce pour proposer, par exemple, des offres de crédits très ciblées et diminuer encore une fois les taux d'attrition. Pour les assurances, il s'agit d'analyser de grands volumes de données pour analyser le comportement des automobilistes grâce aux boites noires qui se multiplient dans les voitures. HP travaille également avec une distribution Hadoop, celle de Cloudera, associée à sa base de données Vertica et à Autonomy IDOL pour fournir des outils d'analyse prédéfinis. Grâce au paquet Hadoop d'Autonomy, les utilisateurs peuvent incorporer un moteur IDOL 10 dans chaque noeud de leur cluster Hadoop. Ce qui leur permet ensuite d'accéder à 500 fonctions d'analyse et de synthèse des données IDOL dans Hadoop.

Terminons notre panorama du salon avec **Bull** qui s'est associé avec Microsoft pour pousser ses solutions big data. Jean François Vannier, responsable commercial infrastructures décisionnelles chez **Bull**, nous a détaillé l'offre Better Data . Elle repose sur la plate-forme datawarehouse de Microsoft, SQL Server FastTrack - une appliance - capable de traiter en temps jusqu'à une centaine de téraoctets. Avec AT Internet par exemple pour du web analytique. Et pour monter en puissance, Bull va bientôt avancer l'offre Parallel Data Warehouse 2.0, une plate-forme capable de supporter jusqu'à 5 Po de données. Elle utilise un moteur, baptisé PolyBase, qui prend en charge des requêtes sur des données relationnelles et non relationnelles avec Apache Hadoop. Les requêtes Hadoop seront acheminées via le logiciel de datawarehouse Apache Hive.

Article de Serge Leblal

Document N°5 Big data, un nouveau défi pour les entreprises

LeMonde Informatique Publié le 24 février 2013 par amecsi_admin

Ce n'est plus une surprise, le concept de big data devrait avoir la part belle des investissements en 2012. Et pour cause, ce phénomène est la réponse à la multiplication des données à traiter, surtout celles qui ne sont pas ou peu structurées en provenance des réseaux sociaux, des sites de vidéos en ligne, des emails, etc. «La consumérisation IT a fortement contribué à l'explosion des données dans les entreprises et donc du big data » indique à juste titre Fabrice Endlicher, directeur des ventes Europe du Sud et Benelux du distributeur à valeur ajoutée Stordis.

En 2011, 1,8 zétabytes de données se sont échangés selon le cabinet d'études américain IDC, ce qui équivaut à remplir 57,5 milliards d'iPad d'une capacité de 32 Go. L'augmentation effrénée des volumes de données est aussi à mettre en parallèle avec l'explosion de la data mobile (usage intensif des smartphones), des usages convergents et multi-terminaux en entreprise. Bref, selon IDC, d'ici à 2020, l'accroissement du volume d'information à gérer sera 50 fois supérieur.

Fort de ce constat, on comprend mieux le phénomène du big data. Mais, cette croissance exponentielle des données interpelle sur la gestion de leur cycle de vie, leur qualité, leur sécurité et surtout leur traitement. Selon une étude menée par l'éditeur LogLogic et le cabinet de conseil Echelon One en janvier-février auprès d'un panel international de 207 entreprises, 67% de ces dernières avouent que la gestion de toutes ces données représente un enjeu important voire sensible. Sensible car les données semi-structurées et non structurées sont problématiques à exploiter contrairement aux informations transactionnelles du système d'information.

Les entreprises sont et seront confrontées aux limites des systèmes existants de base de données relationnelles qui ne sont plus à même de traiter et d'être ainsi rentables pour faire face à ces volumes énormes de données non structurées. « Il y a 10 ans, les données étaient à 70% structurées pour 30% d'informations non structurées. Demain ça serait l'inverse. Cela signifie que le datawarehouse risque de s'isoler de plus des 2/3 des données. Ce n'est donc plus acceptable » s'alarme Gilles André, directeur général de Polyspot. De plus, le poids de la donnée non structurée est extrêmement important ce qui dégrade les performances. Par exemple, l'exécution de certaines ressources et requêtes (plusieurs milliards de lignes) critiques pour l'activité de l'entreprise peut

réclamer plusieurs heures à partir des systèmes traditionnels alors qu'il faut au mieux quelques secondes sur des plateformes nouvelles générations très évolutives de type «big data ».

Parallèlement à la vétusté des systèmes déjà existants en interne, les entreprises font aussi face à un déficit en compétences nécessaires pour exploiter les possibilités qu'offre le croisement des big data avec l'analyse de données. Il faut dire que les opérations à réaliser (chargement de données, extraction, transformation, traitement, etc.) réclament une certaine expertise dans ce domaine. D'ailleurs, selon les résultats d'une enquête effectuée par le groupe EMC auprès d'environ 500 data scientists, un spécialiste du big data, au niveau mondial, la majeure partie des entreprises aujourd'hui souffrent de cette carence en compétences. Ainsi, 32% des personnes interrogées estiment qu'elles manquent de compétences et de formation, 32% d'entre elles confient un manque de budget ou de ressources, 14% des entreprises avouent qu'elles n'ont pas une structure organisationnelle adaptée et 10% des personnes relèvent une pénurie d'outils et de technologies.

Un nouvel eldorado pour les éditeurs

Globalement, le big bata reprend les mêmes problématiques que le décisionnel (Business Intelligence) même si, théoriquement, la BI se concentre davantage à l'exploitation de données à des fins de monitoring et d'aide à la décision, alors que le big data, s'ouvrant en plus à des informations présentes à l'extérieur de l'entreprise, est plus focalisé sur l'utilisation d'outils d'analyse et d'algorithmes afin de générer des données prospectives devant mener à l'innovation. Cela dit, big data ou Business Intelligence, tous deux ont pour objectif des analyses de données pertinentes et en temps réel de préférence.

Ainsi, tous les acteurs spécialisés dans l'analytique, l'intégration, l'extraction et l'analyse de données s'intéressent à ce nouvel eldorado pour pousser leurs offres. On peut citer Japersoft, Pentaho ou Quantum 4D pour leur solution décisionnel mais aussi Terradata avec Aster Data Analytic Platform, EMC avec Greenplum Data Appliance Computing, Oracle avec Exadata, HP avec Vertica Analytics Platform, IBM avec Netezza et BigInsights, SAS avec Business Analytics, SAP avec Sybase IQ VLDP, Talend Open Studio for big data, etc.

Document N°6 Edicia associe big data et sécurité urbaine

LeMonde Informatique - article de Maryse Gros 19/02/2014

Le portail de sécurité d'Edicia renseigne le maire d'une ville en temps réel sur les événements de sécurité enregistrés dans la journée. (cliquer sur l'image / chiffres fictifs)

Editeur d'un SI destiné aux polices urbaines, Edicia complète sa solution d'indicateurs de sécurité pour informer les maires en temps réel sur les éléments recueillis dans la journée. Les données remontées d'Internet sont traitées dans un évaluateur de risques développé par un laboratoire de recherche. Des apps mobiles permettent une interactivité avec les citoyens.

Spécialisé sur les solutions de sécurité urbaine, l'éditeur de logiciels Edicia a développé un système d'information qui s'adresse principalement aux polices municipales, mais qui a vocation à s'étendre plus largement, notamment vers les établissements recevant du public (banques, centres commerciaux...). La solution s'enrichit maintenant d'un portail de sécurité qui va permettre au maire, au directeur général des services ou au responsable du centre de supervision urbain de consulter des indicateurs en temps réel sur les différents aspects de sécurité dans la ville : prévention des risques, gestion du stationnement, dysfonctionnements divers, etc.

« Alors que le logiciel destiné aux directeurs de police est vraiment très métier, le tableau de bord va cette fois permettre à différentes personnes, en fonction de leurs prérogatives, de comprendre rapidement une situation, par exemple pour répondre à une demande du préfet », explique Annie Bourget, directrice marketing d'Edicia. La solution dispose de 80 indicateurs environ. « Nous préconisons d'en afficher huit, les essentiels, en fonction des attributions de la personne qui les consultera ». Les informations sont remontées du terrain dans le SI par des agents équipés de terminaux mobiles. L'application leur permet de rédiger des rapports, saisir des mains courantes ou verbaliser. Développée en HTML5, elle peut être déployée sur Android, iOS ou Windows Phone. Edicia propose de son côté le terminal durci TC55 de Motorola.

Big data : évaluer les risques à partir des rumeurs remontées du web

L'un des intérêts de la solution d'Edicia, déjà installée dans 470 villes en France, c'est qu'elle propose aussi de récupérer les données transmises par les citoyens via une app mobile, Risk. Celle-ci permettra de transmettre des dysfonctionnements, par exemple des dégradations sur les équipements municipaux ou des problèmes d'inondation. Ces informations seront raccrochées au SI global pour en renforcer la pertinence. Une app

de ce type est notamment proposée depuis deux ans dans la ville de Nice.

Par ailleurs, la solution prend maintenant en compte la dimension big data. « Notre outil permet de remonter les données extraites de rumeurs sur les réseaux sociaux, de qualifier ces données et d'adresser ensuite des moyens, l'outil gérant aussi la disponibilité des agents », indique Annie Bourget. « Nous traitons ces données dans un évaluateur de risques. Par exemple, s'il y a une rumeur d'apéritif géant, il pourra réunir 10 000 personnes s'il fait beau et seulement 2 000 en cas de mauvais temps. S'ils ont l'information, les responsables de la sécurité vont pouvoir mettre en place les moyens adéquats ». Edicia travaille avec le laboratoire d'informatique de l'Université de Nantes Atlantique, le Lina. « Ils ont développé une expertise dans la modélisation du risque avec un outil maison qu'ils appliquent à différents domaines ». Dans le cas de la solution Edicia, elle est appliquée à la sécurité urbaine, en tenant compte de l'expérience des responsables de sécurité, dont la police des transports (SNCF, RATP, régie de transport d'une ville).

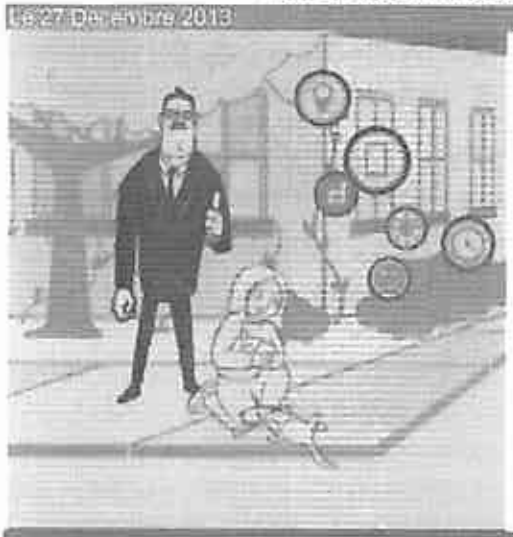
Des alertes en push sur l'app mobile

L'application Risk sur smartphone peut aussi permettre à l'abonné de s'abonner aux alertes de son choix, par exemple sur les intempéries (montée d'un cours d'eau), ou sur les événements exceptionnels, organisation de manifestations (pour les habitants du périmètre). « Le type d'alerte est défini par la ville, tout est paramétrable », précise Annie Bourget. « Nous travaillons avec le Québec qui avertit du passage de la déneigeuse pour que les habitants puissent déplacer leur véhicule ». Le 3ème volet de l'app mobile, c'est la diffusion d'information pour améliorer la culture du risque chez les usagers, « pour qu'ils sachent qu'en cas de problème, ils peuvent écouter la radio, par exemple », ajoute la directrice marketing. Les « push alertes » peuvent se faire par message vocal, par SMS ou par push sur l'app mobile. « En multipliant les canaux, il n'y a pas le même encombrement en situation de crise », souligne-t-elle.

Document N°7 Le big data au cœur du rapport 5in5 d'IBM

LeMonde Informatique un article de Jacques Cheminat

Le 27 Décembre 2013



IBM voit dans les 5 prochaines années la création d'un ange gardien numérique. Crédit Photo: IBM

Dans son rapport annuel détaillant 5 innovations pour les 5 prochaines années, IBM a focalisé ses travaux sur le big data avec une application dans l'enseignement, la médecine, le commerce, la ville et la sécurité.

Chaque année, **IBM** publie **un rapport baptisé 5in5** qui donne les tendances des innovations sur les 5 prochaines années. Cette édition est clairement orientée vers le big data et sur l'usage fait de la collecte et du traitement de cet afflux massif de données. Le rapport distingue 5 scénarios impliquant le big data.

Le premier concerne l'école. A partir de différentes informations collectées sur les élèves, parcours, résultats, les cours s'adapteront automatiquement au niveau de l'enfant. Les données permettent aussi aux professeurs de repérer les élèves en difficulté. Les exercices seront ajustés en fonction de l'intérêt de l'enfant pour certaines matières. Des tests ont été réalisés par IBM avec l'école publique de Gwinett dans l'Etat de Georgie.

Le second s'applique au commerce. Les boutiques devraient en savoir de plus en plus sur les habitudes des consommateurs et amélioreront leurs offres en fonction. IBM utilise son supercalculateur Watson pour analyser les comportements des consommateurs pour mieux cibler leurs attentes. Ainsi, les laboratoires de Big Blue travaillent sur un prototype de logiciel, Virtual Stylist, qui à partir des données collectées donnent aux confectionneurs d'habits les souhaits des clients en matière de vêtements. Aujourd'hui, cette prescription se fait par rapport à vos achats précédents.

Le troisième scénario parle de médecine. IBM veut là aussi s'appuyer sur Watson dans la recherche des traitements contre le cancer. Cette recherche se veut transversale, mêlant des données médicales (recherches, expérimentations, thérapies, publications) et les données des patients (résistance au traitement, effets secondaires, analyses). L'idée est de mélanger l'ensemble de ces informations dans le cloud, de mettre à disposition des médecins des options de protocoles et de choisir le meilleur traitement avec le patient. Par ailleurs, avec les travaux sur l'ADN, il sera possible avec les capacités d'analyse des big data d'aller vers des traitements personnalisés au niveau génomique.

Ville connectée et ange gardien numérique

Le quatrième modèle met en scène la ville. Elle devient de plus en plus connectée, à travers les capteurs, les applications mobiles, les données partagées (crowdsourcing), les réseaux sociaux, etc. L'ensemble des informations donnent aux collectivités locales la possibilité de prévoir et d'interagir avec leurs concitoyens. IBM souligne qu' « en 2017, le nombre de smartphones dans le monde dépassera trois milliards. Les gens auront une clé numérique de la ville au bout des doigts ». Des initiatives sont déjà en action. Ainsi, à Toulouse, la mairie a travaillé pour cerner les signaux faibles et s'adapter. Par exemple, les signalements d'encombrants, de travaux, d'absence d'éclairage dans la rue, etc. La municipalité peut mieux gérer ses équipes pour intervenir.

Enfin, le dernier scénario est la création d'un ange gardien numérique. Il ne se passe pas une journée sans des incidents de sécurité informatique impliquant des vols de données (mots de passe, identifiants, coordonnées bancaires, etc). Ce gardien va être capable d'analyser un grand volume de données pour diagnostiquer votre comportement en ligne et ainsi détecter des risques de vols ou d'actions anormales. En utilisant l'informatique cognitive, l'ange gardien sera capable de recouper des informations et de trouver des contradictions, comme acheter un hot dog dans une station-service, alors que vous êtes végétariens ou payer de l'essence alors que votre réservoir est plein.

Document N°8 Panorama des solutions de big data

Ces bonnes feuilles sont extraites de l'ouvrage "Enjeux et usages du Big Data" (Collection Management et informatique), de C. Brasseur, chez ©Lavoisier, 2013.

Panorama des solutions de big data : des solutions majoritairement open source

Sommaire

Les grands acteurs du web tels que Google, Yahoo ou Facebook ont été les premiers à être confrontés à des volumétries de données extrêmement importantes, et les principales innovations se retrouvent sans surprise parmi ces pionniers. Les développements portent essentiellement sur deux types de technologies :

les bases de données ;

les plates-formes de développement et de traitement des données.

En savoir plus

- **Ouvrage :** Enjeux et usages du Big Data

Ces entreprises innovantes ont choisi pour la plupart d'ouvrir le code initialement développé en interne pour en faire des projets open source. Le tableau 4.3 présente quelques exemples de technologies open source utilisées pour la gestion des données massives et dont l'origine est un développement interne

Un certain nombre des technologies citées ci-après comme Hadoop et Cassandra font partie de la fondation Apache, organisation à but non lucratif qui développe des logiciels open source, dont le célèbre serveur Apache HTTP Server. Les objectifs principaux de la fondation sont de protéger juridiquement le travail des contributeurs et d'empêcher que la marque Apache soit utilisée illégalement.

Il est intéressant de souligner que les grands acteurs du logiciel ont complètement intégré la dimension open source, en proposant dans leurs offres dédiées au big data des briques basées sur ces technologies. Ainsi, Oracle a mis Hadoop au cœur de son offre "Big Data Appliance",Microsoft a également intégré Hadoop au sein de son offre Windows Azure,de même que IBM, EMC et Netapp pour leurs offres de gestion de données volumineuses. Parmi les différentes technologies développées, Hadoop apparaît clairement comme une solution de référence.

Société	Technologie développée	Type de technologie
Google	Big Table	Système de base de données distribuée propriétaire reposant sur GFS (*Google File System*). Technologie non *open source*, mais qui a inspiré HBase qui est *open source*
	MapReduce	Plate-forme de développement pour traitements distribués
Yahoo	Hadoop	Plate-forme Java destinée aux applications distribuées et à la gestion intensive des données. Issue à l'origine de Google BigTable, MapReduce et Google File System
	S4	Plate-forme de développement dédiée aux applications de traitement continu des flux de données
Facebook	Cassandra	Base de données de type NoSQL et distribuée
	Hive	Logiciel d'analyse de données utilisant Hadoop
Twitter	Storm	Plate-forme de traitement de données massives
	FlockDB	Base de données distribuée de type graphe
LinkedIn	Kafka	Système distribué de gestion des messages
	SenseiDB	Base de données temps réel distribuée et semi-structurée
	Voldemort	Base de données distribuée destinée aux très grosses volumétries

Tableau 4.3. Quelques technologies open source du big data. © Lavoisier

Document N°9 Relation client, Auchan mise sur Proxem pour analyser ses big data

Lemonde Informatique article de Bertrand Lemaire 18/2/2014

Le 18 Février 2014

La solution de Proxem permet à Auchan d'identifier les problèmes pointés du doigt par les consommateurs et d'analyser le succès ou non d'un produit

Pour exploiter au mieux les différents commentaires de ses clients, Auchan s'est doté d'une solution d'analyse sémantique big data multicanale.

Comme beaucoup d'entreprises aujourd'hui, Auchan cherche à mieux connaître ses clients pour améliorer la qualité de ses relations avec eux. L'enseigne française de grande distribution veut identifier précisément leurs profils, leurs attentes, les problèmes qu'ils rencontrent ainsi que leurs avis sur les produits, l'enseigne et ses concurrents. Il fallait trouver un moyen d'analyser en temps réel le ressenti de ses clients à travers les études marketing, les mails qu'ils lui envoient mais aussi les commentaires qu'ils postent sur internet, notamment les réseaux sociaux. L'enseigne a finalement choisi la solution d'analyse sémantique big data multicanale de Proxem, Ubiq Voix du Client.

Concrètement, la solution de la société française spécialisée dans l'analyse big ata permet à Auchan, d'identifier les problèmes pointés du doigt par les consommateurs, d'analyser le succès ou non d'un produit, de détecter les différences de prix avec ses concurrents et bien évidemment, de mesurer l'évolution de la satisfaction client. Grâce à ces différentes données et leur distribution personnalisée entre les différents échelons de l'entreprise, Auchan est ainsi en mesure d'optimiser sa politique de prix et d'anticiper les demandes des consommateurs, les problèmes de qualité et les ruptures de stock. Le montant du projet n'a pas été dévoilé.

Document N°10 — Sans gouvernance, une perte de contrôle probable des données

LeMonde Informatique article de Benoit Huet – février 2014

Sans gouvernance, une perte de contrôle probable des données

À en croire nos interlocuteurs, une majorité d'entreprises manquent clairement de visibilité sur la gestion de leurs données. « *Le constat est clair, il manque dans les entreprises une gouvernance et un réel manque de conviction pour la gestion des données. Par exemple, concernant l'archivage, nombreuses sont les entreprises à conserver des données pendant 20 à 30 ans alors qu'elles n'ont plus aucune valeur légale ! De plus, le problème n'est pas uniquement lié à la quantité des données, mais aussi à la qualité de la donnée. Celle-ci doit être, aujourd'hui, utilisée à bon escient. Si j'ai multiplié par deux ma capacité de stockage, mais, en parallèle, j'ai multiplié par quatre la qualité, où est le problème ? En revanche, si je multiplie ma capacité et je n'en tire aucune valeur, cela pose problème* », explique Jean-Baptiste Ceccaldi, PDG et fondateur de Sentelis. Selon Steria, par manque de gouvernance, 40 % des coûts informatiques d'entreprise sont imputables à des problèmes de qualité des données, tandis que 30 % des données conservées par les entreprises s'avèrent incomplètes et erronées.

Dépassées, les entreprises ont besoin d'aide...

La banalisation généralisée de la gouvernance des données n'existe pas dans toutes les entreprises, loin de là (hormis quelques grands comptes comme Meteo France), il y a plutôt un manque réel de gouvernance. Ce dernier n'est d'ailleurs pas toujours dû à une négligence forcée de la part des entreprises, elles sont bien conscientes qu'elles ont besoin de visibilité et les sondages en attestent. De plus comme le rappelle Cyril Van Agt, responsable avant-vente partenaires chez NetApp France, la gestion du stockage est souvent perçue comme la cinquième roue du carrosse.

Dans ce contexte, le problème devient de plus en plus difficile à gérer pour une entreprise. D'une part, elles font face à une croissance exponentielle du volume des données et, d'autre part, à une intensification des réglementations (Sarbanes Oxley, Bale 3 pour les banques, Solvency pour les assurances, etc, sans oublier la réglementation sur la protection des données à caractère privé et personnel). D'où l'importance du rôle des prestataires qui interviennent auprès des entreprises. « *Les SSII doivent, plus encore aujourd'hui, mettre en place un plan de trafic pour faire respecter la gestion et le comportement du système en fonction des attributs des données et faire de*

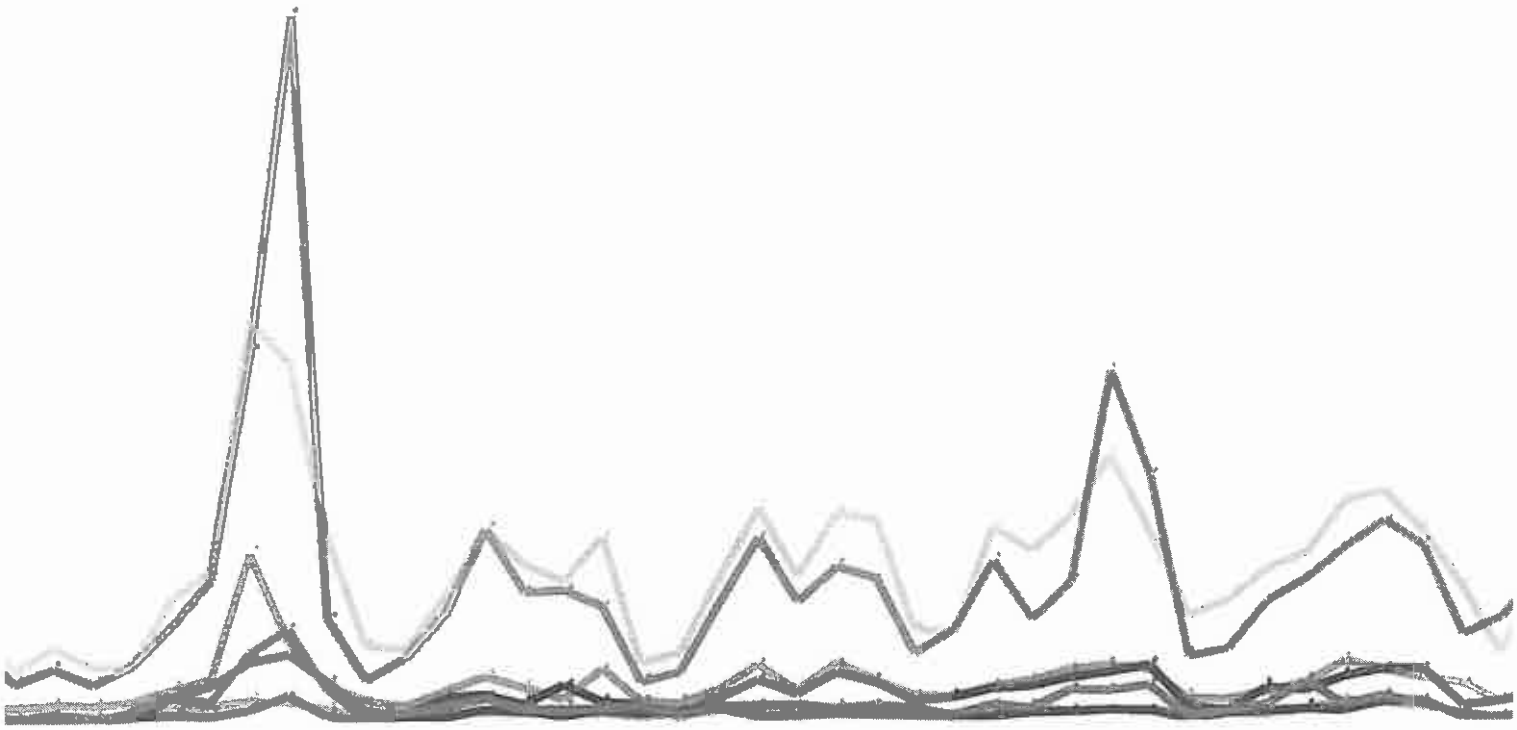
l'analyse préventive », insiste Vincent Videlaine, directeur EMEA Strategic Alliances et Services Providers chez Symantec. Car si les entreprises ne restent pas inactives face à la gestion de leurs données, elles sont plus dans une approche palliative et moins curative.

Une gouvernance plutôt en phase d'expérimentation « *Aujourd'hui c'est vrai, sur ce sujet complexe, nous sommes plus dans un mode d'expérimentation de la gouvernance, il n'y a pas encore un phénomène d'industrialisation* », reconnaît Jean-Baptiste Ceccaldi. Et d'ajouter : « *le degré de maturité est faible, car les solutions ne répondent pas forcément aux enjeux métiers, il n'existe pas une seule gouvernance, mais plusieurs en fonction des métiers, ce qui marche pour une entreprise ne marche pas forcément pour l'autre et c'est aussi valable pour les services en interne* ». Pour répondre aux métiers, les fournisseurs de solutions et de services de stockage se doivent donc d'être plus attentifs à leurs besoins réels, ceci est valable à la fois pour les grands comptes, mais encore plus pour les petites et moyennes entreprises qui ne disposent pas forcément de compétences en interne. « *Pour notre coeur de cible, à savoir les PME et les TPE, nous essayons d'avoir des approches métiers, car les contraintes ne sont clairement pas les mêmes. Pour les grands comptes, je ressens effectivement que les DSI ont eu une connaissance insuffisante des métiers, il manque une couche fonctionnelle* », remarque Luc D'Urso, PDG de Wooxo.

De plus, la gouvernance au niveau des métiers implique aussi pour les entreprises un travail collaboratif indispensable. Les équipes de différents services se doivent de travailler ensemble sur les problèmes de gestion de données ce qui est loin d'être le cas... Pour Guy Chesnot, architecte stockage chez SGI, ce manque de communication s'explique aussi par le manque de compétences en interne. « *Il y a encore très peu de datascientists bien formés dans les entreprises, mais cela change progressivement avec la médiatisation autour des big data, les écoles et les universités en parlent de plus en plus et forment de futurs candidats...* »

Quant aux entreprises qui sont plutôt dans une attitude « d'attentistes » face à cette gouvernance en se disant que des solutions technologiques à venir les feront à leur place, c'est pour le dirigeant du cabinet Sentelis une illusion qui n'empêchera pas la prolifération des données. Au final, les entreprises qui mènent à la fois la maîtrise de l'accroissement de la donnée et de sa qualité seront les grandes gagnantes à en croire les personnes que nous avons interviewées.

DOCUMENT n° 11



Dan Jewett, Vice-président, service Gestion des produits

7 conseils pour maîtriser les Big Data en 2014

Alors que l'expansion des Big Data pouvait vous sembler stabilisée, celle-ci se poursuit encore. Indépendamment de leur volume réel, les Big Data démontrent actuellement leur valeur. Partout dans le monde, diverses organisations disposent de telles informations qui présentent des structures et des volumes variés. Ces organisations comprennent l'importance et la pertinence de ces données, et se sentent obligées de leur porter une attention particulière. Il est désormais évident que les Big Data survivront à ceux qui les ignorent.

Les organisations qui ont déjà apprivoisé les Big Data, c'est-à-dire la masse multi-structurée qu'elles avaient stockée avant d'en connaître la valeur, optimisent leur efficacité opérationnelle, la croissance de leurs revenus et la mise en place de leurs nouveaux modèles d'affaires.

Comment procèdent-elles ? Leurs techniques fructueuses peuvent être résumées en sept conseils.

1. Prévoir à long terme en réfléchissant à court terme.
2. Détecter les fausses interrogations.
3. Présenter les Big Data sous forme visuelle.
4. Permettre aux utilisateurs d'obtenir des informations essentielles.
5. Générer un grand volume de données à partir de sources restreintes.
6. Veiller à ce que les Big Data ne créent pas de graves problèmes.
7. Mettre à exécution.

1.

Prévoir à long terme en réfléchissant à court terme

Vous n'êtes pas seul à vous préoccuper de l'évolution des technologies liées aux Big Data. Leurs caractéristiques évoluent si rapidement qu'il est impossible de connaître les outils, les plates-formes et les méthodologies qui seront à l'œuvre durant les deux prochaines années.

Détendez-vous. Votre activité peut parfaitement s'adapter à cette évolution rapide.

Chaque année, les fournisseurs améliorent leur utilisation des Big Data. Les systèmes transactionnels relationnels en ligne (OLTP) deviennent plus efficaces et intelligents, qu'ils soient exécutés sur site ou dans le cloud. Les derniers développements techniques facilitent l'interfaçage entre Hadoop et les entrepôts de données. Et de nouveaux produits apparaissent sans cesse sur le marché, répondant toujours plus précisément à vos besoins.

Inutile donc de vous inquiéter. Restez à l'écoute des possibilités offertes par ces nouveaux produits, aussi longtemps qu'ils fournissent assez de valeur pour justifier leur intégration à votre environnement actuel. Maintenez une plate-forme d'analyse décisionnelle capable d'interagir directement avec de nombreux formats différents. Vous serez alors prêt à accueillir tout ce que le marché peut produire

2.

Détecter les fausses interrogations

Votre organisation a-t-elle besoin d'Hadoop ou d'un entrepôt de données ? Cette interrogation n'est qu'une question-piège. Non seulement Hadoop et les entrepôts de données fonctionnent parfaitement en parallèle, mais les organisations profitent même des capacités de collaboration de ces systèmes.

Un entrepôt de données permet de mieux traiter vos données importantes et structurées, puis de les stocker afin que vos outils d'analyse décisionnelle et vos tableaux de bord puissent y faire appel facilement. Mais cette ressource sera moins efficace et plus lente dans le cadre du traitement analytique et de certains autres types de transformations.

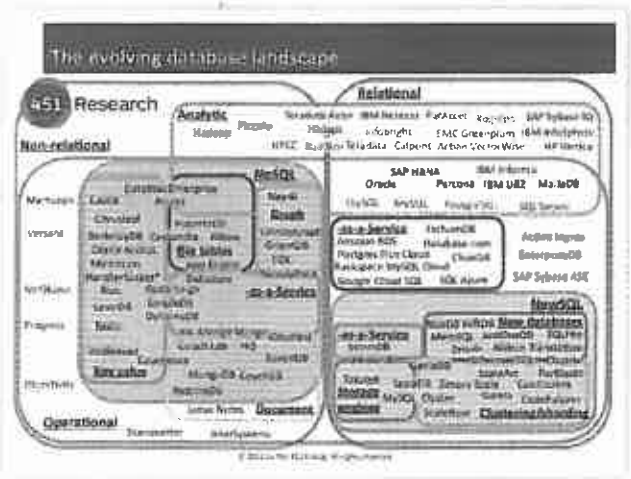
Laissez donc Hadoop se charger de ces tâches. De plus, bien qu'Hadoop soit peu adapté aux requêtes interactives et à la gestion des données, il excelle dans l'intégration des données brutes, non structurées et complexes.

Ensemble, ces systèmes fonctionnent en symbiose. Pensez par exemple aux données sur lesquelles s'appuient vos dirigeants lorsqu'ils anticipent leurs besoins matériels pour l'année à venir. Ce jeu de données est probablement gigantesque et vous n'avez pas assez de temps pour le modéliser, le structurer ou le préparer d'une manière ou d'une autre à son intégration dans votre entrepôt de données. Lorsque les dirigeants concernés auront fini de le traiter, parfois en une semaine seulement, ils souhaiteront s'en débarrasser. C'est à ce moment précis qu'Hadoop intervient pour stocker et raffiner ces données, avant d'en transmettre un échantillon à l'entrepôt de données.

« Les Big Data ne remplacent pas le stockage en entrepôt de données », a écrit Mark Madsen, PDG de Third Nature, dans son article « What big data is Really About ». « Elles ne constituent pas non plus un îlôt à maintenir séparément. Elles font partie du nouvel environnement informatique. »

Veillez à ne pas tomber dans le piège « Hadoop ou entrepôt de données ? ». Vous pouvez et devez utiliser ces deux ressources.

Simplifier et faire coexister



Source : Matthew Aslett, The 451 Group
Updated database landscape graphic, 2 nov. 2012

Requirement	Data Warehouse	Hadoop
Low latency interactive reports and OLAP	•	
ANSI 2003 SQL compliance is required	•	
Preprocessing or exploration of raw unstructured data		•
Online archives alternative to tape		•
High-quality cleansed and consistent data	•	
100s to 1000s of concurrent users	•	•
Discover unknown relationships in the data	•	•
Parallel complex process logic		•
CPU intense analysis	•	•
System, users, and data governance	•	
Many flexible programming languages running in parallel		•
Unrestricted, ungoverned sand box explorations		•
Analysis of provisional data		•
Extensive security and regulatory compliance	•	
Real time data loading and 1 second tactical queries	•	•

Source : Dr Amr Awadallah et Dan Graham « Hadoop and the Data Warehouse: When to Use Which » conjoint par Cloudera, Inc. et Teradata Corporation, 2012

« Une visualisation habile et réfléchie permet d'avoir des traits de génie. Il est tout simplement impossible d'obtenir le même résultat avec une feuille de calcul. »

— Dana Zuber, Wells Fargo

3.

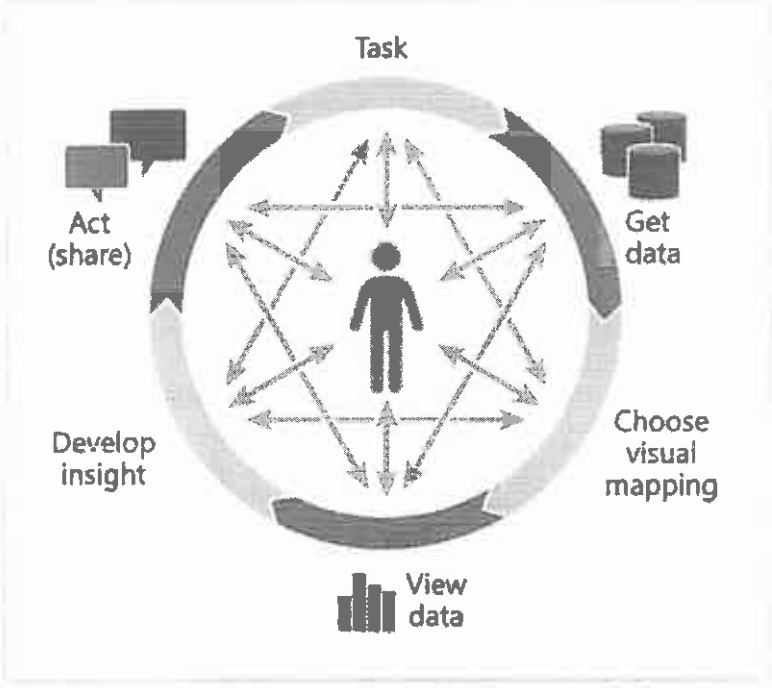
Présenter les Big Data sous forme visuelle

Les Big Data deviennent visuelles lorsque vous pouvez les afficher graphiquement. Un rapport établit en 2013 par Aberdeen Group a révélé qu'« au sein des organisations qui utilisent des outils d'exploration visuelle, 48 % des personnes recourant à l'analyse décisionnelle sont capables de trouver l'information par elles-mêmes, sans demander l'intervention de l'équipe informatique ». Sans recours à l'exploration visuelle, ce taux chute à seulement 23 %.

Cette étude montre également que les gestionnaires faisant usage de l'exploration visuelle des données ont 28 % de chance de plus de trouver des informations à jour, par rapport à leurs homologues dépourvus d'outils visuels.

Aspect sans doute capital en matière de Big Data, le rapport précise que la visualisation encourage aussi l'interaction avec les données. Les gestionnaires manipulant des données visuelles ont plus de deux fois plus de chance (33 % contre 15 %) d'interagir en profondeur avec celles-ci. Ils sont en outre plus susceptibles de poser des questions à la volée, souvent inspirées par des informations aperçues quelques instants auparavant.[1]

Explorer visuellement les données, c'est leur permettre de s'exprimer pleinement, d'une manière que le cerveau peut comprendre instantanément. « L'effet est comparable à celui d'un coup de génie », a commenté Dana Zuber, vice-présidente de la planification



Source : Dr Jock D. Mackinlay « How to See and Understand big data » 2007

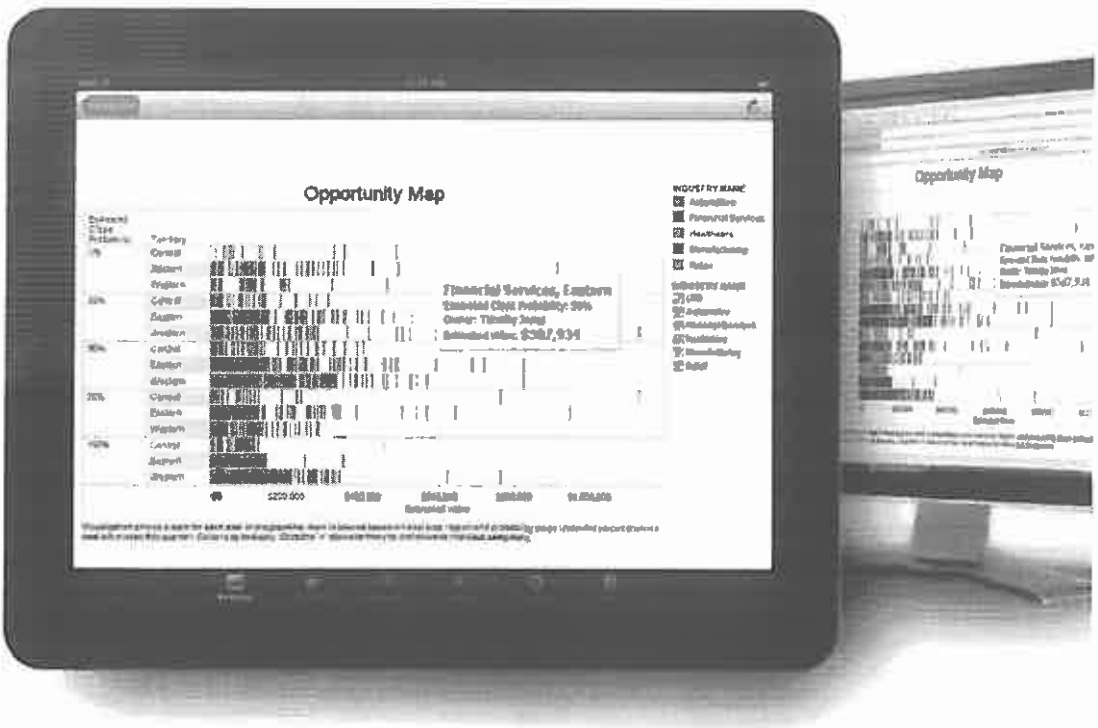
stratégique, Wells Fargo. « Il est tout simplement impossible d'obtenir le même résultat avec une feuille de calcul. »

L'analyse visuelle vous permet d'accomplir deux actions à tout moment :

- Modifier les données que vous visualisez,
  car des questions différentes nécessitent souvent des données distinctes.
- Modifier la façon dont vous les visualisez,
  car chaque vue peut répondre à plusieurs questions.

Grâce à ces étapes simples, vous enclenchez ce que l'on appelle le « cycle de l'analyse visuelle » : obtenir des données, visualiser ces données, poser des questions et obtenir des réponses, et ainsi de suite. À chaque cycle, vos demandes se précisent de plus en plus, proportionnellement aux informations que vous obtenez. Vous pouvez explorer les données en profondeur, en surface ou latéralement. Vous pouvez également faire surgir de nouvelles données. Il devient possible de créer différentes vues successives, à mesure que vos visualisations accélèrent et développent votre réflexion.

Lorsque tout est prêt, vous n'avez plus qu'à partager. Vos collaborateurs posent leurs propres questions et trouvent les réponses adaptées, accélérant ainsi la perception, l'activité et la productivité de votre équipe.



Consultez et manipulez le tableau de bord Web en direct [illegible] sur votre tablette Android

4.

Permettre aux utilisateurs d'obtenir des informations essentielles

Avez-vous déjà rencontré des personnes passionnées par l'information ? Rien ne peut les arrêter. Elles posent sans cesse de nouvelles questions et génèrent de la valeur jusqu'à être convaincues de tout maîtriser... ou jusqu'au moment où elles doivent demander de l'aide au service informatique.

Avec les Big Data, cette passion pour l'information s'exprime encore plus ardamment. Jusqu'à devenir brûlante. Le temps est désormais entièrement dévolu à l'analyse de données en libre accès.

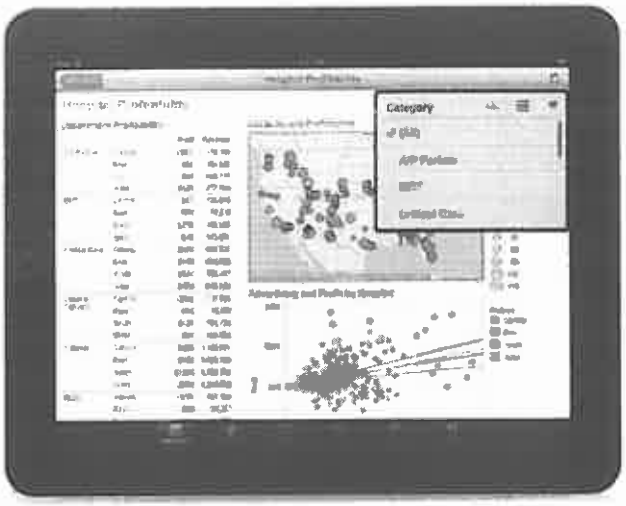
« Les organisations utilisant les Big Data ont plus de 70 % de chance de plus que les autres de voir leurs projets d'analyse décisionnelle menés essentiellement par leurs analystes professionnels, plutôt que par leur service informatique », a déclaré Aberdeen Group dans son récent rapport « Go Big or Go Home? Maximizing the Value of Analytics and big data ».

Avec les Big Data, les utilisateurs professionnels ne tolèrent pas l'obsolescence et la lenteur de certaines méthodologies informatiques qui publient les données comme les chapitres d'un livre.

Dans son étude « The Value of big data », Mark Madsen, analyste de recherche chez Third Nature, a précisé : « Le modèle de publication de l'analyse décisionnelle est archaïque, tout comme le contexte d'utilisation des informations qu'il instaure. Cette situation est comparable à la lecture d'un livre à la lueur d'une lampe torche ou d'une chandelle : il s'agit de ce que l'on appellait des "élucubrations". »[2]

« Les Big Data agissent au même titre qu'un éclairage électrique moderne », a ajouté M. Madsen. « Elles illuminent les zones qui restaient auparavant obscures. Elles fournissent un meilleur éclairage, ainsi que la capacité d'en disposer dès que nécessaire. Plutôt que d'attendre plusieurs mois que les données soient parfaitement nettoyées et prêtes à l'emploi, il devient possible d'utiliser les technologies liées aux Big Data pour rechercher et découvrir la valeur qu'elles recèlent. Lorsqu'elles s'avèrent valables, les données peuvent être traitées par les processus les plus stricts afin d'être stockées dans un entrepôt de données. »

N'obligez pas les utilisateurs à « élucubrer ». Donnez-leur une liberté d'investigation totale.



Source : Tableau Software
Solution décisionnelle mobile

5.

« Il est important de fusionner toutes ces données pour comprendre les raisons qui poussent les consommateurs à entrer dans une boutique et à remplir leur panier. »

— Rishi Kumar, Unilever

[c1] *En savoir plus sur le point de vue de M. Kumar au sujet de la fusion des données.*

Générer un grand volume de données à partir de sources restreintes

En regardant attentivement, vous découvrirez les constituants des Big Data : d'innombrables jeux de données de plus petite taille. Utilisé séparément, chacun de ces jeux de données peut fournir de la valeur. Une fois réunis, ils représentent une valeur inestimable.

Dans le secteur des biens de consommation, par exemple, les dirigeants ne peuvent comprendre pleinement le comportement des consommateurs qu'une fois les données psychologiques fusionnées avec les données commerciales.

« Les cartes de fidélité fournissent des données d'une grande richesse », a expliqué Rishi Kumar, directeur des analyses, Unilever. « Il est important de fusionner toutes ces données pour comprendre les raisons qui poussent les consommateurs à entrer dans une boutique et à remplir leur panier. » Cette approche permet à Unilever d'anticiper la popularité des produits et les tendances émergentes.

Une part importante de cette valeur revient aux organisations qui réunissent leurs données relationnelles, semi-structurées et brutes, avec un minimum d'investissement préalable et sans imposer de contrainte technologique aux utilisateurs professionnels. Le travail peut être accompli, et c'est bien suffisant.

Que vos données se trouvent dans une feuille de calcul, une base de données, un entrepôt de données, des systèmes de fichiers open source tels que Hadoop, ou toutes ces possibilités à la fois, vous devez pouvoir vous connecter rapidement aux données et les consolider.

Vous pouvez ainsi poser des questions et y répondre dès qu'elles se présentent, quel que soit le volume de vos données.

6.

Veiller à ce que les Big Data ne créent pas de graves problèmes

Les Big Data sont aussi amusantes qu'un bac à sable. Vous pouvez y accéder, bâtir et structurer des analyses et même prélever du sable pour le glisser dans le pantalon de votre meilleur ami. Vous voyez ? Bien sûr, uniquement sous la supervision d'un adulte.

Cette masse de données recèle de la valeur, notamment parce qu'elle concerne des personnes réelles. Sans entamer de débat sur l'éthique, il faut savoir que les gouvernements estiment qu'elle incite chacun à « bien se tenir ».

Plus de 80 pays disposent aujourd'hui de lois sur la confidentialité des données. L'Union européenne a défini sept « principes de la sphère de sécurité » afin de protéger les données personnelles de ses citoyens. À Singapour, la loi sur la protection des données personnelles est entrée en vigueur en janvier 2013. Aux États-Unis, Sarbanes-Oxley a averti l'ensemble des sociétés cotées en bourse, tandis que le Health Insurance Portability and Accountability Act (HIPAA) définit des normes nationales relatives à la confidentialité des soins de santé.

En conséquence, avant de plonger dans l'océan des Big Data, assurez-vous que vos besoins soient conformes aux normes de gouvernance et de confidentialité. Êtes-vous un organisme de santé soumis aux normes de l'HIPAA ? Exercez-vous une activité dans certaines régions du monde ? Ou comprenez-vous simplement qu'il est sage de prendre des précautions concernant les éléments clés de vos Big Data ?

Dans ce cas, si votre organisation doit assurer sa conformité, une solution évidente est la gestion des données maîtresses, qui encadre l'utilisation des données en son sein. Lorsqu'elle est instituée, vous êtes prêt à agir. Cependant, parvenir à un accord portant sur les définitions et les règles professionnelles peut être lent et complexe pour une organisation.

Complexe, certes, mais certainement pragmatique. N'outrepassez pas les règles de gouvernance dans le but de gagner en agilité et d'obtenir des résultats plus rapidement. C'est ce que recommande Forrester Consulting dans son rapport 2013 intitulé « big data Needs Agile Information And Integration Governance ». Les résultats issus des Big Data nécessitent une gouvernance.

Forrester met en garde contre l'adhésion à « un ensemble unique de normes, de politiques et de pratiques », qui « bride la valeur pouvant être extraite des investissements dans les Big Data et des informations résultantes ».

Au lieu de cela, ce rapport suggère d'adopter une gouvernance adaptée aux capacités d'analyses et aux objectifs, d'établir des « zones » de gouvernance en fonction de la source et du type des données et de réaliser des essais avant de mettre ces règles en production.[3]

7.

Mettre à exécution

Ce dernier conseil est peut-être le plus important de tous : lancez-vous. Plongez et suivez les six étapes décrites précédemment.

Les Big Data sont déjà sur le seuil de votre organisation, voire à l'intérieur de celle-ci. Obtenez des résultats dès maintenant.

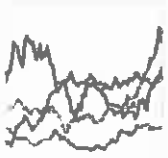
« Je peux répondre au cours d'une réunion, aussi vite qu'en ce moment même. », a affirmé Peter Gilks, Barclays. « Auparavant, nous devions prévoir une attente d'un jour ou deux pour chaque question. Désormais, je peux participer à une réunion avec mon ordinateur portable et répondre à la volée à des questions portant sur 20 millions de lignes. » (Voir les études de cas concernant votre secteur.)

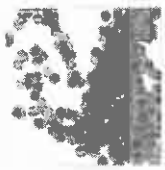
Lorsque vous avez quelque chose à présenter, les autres le remarqueront, car rien n'attire l'attention autant que des résultats. Un cercle vertueux s'enclenche alors et permet de diffuser des résultats dans l'ensemble de l'organisation.

Un dirigeant pourra alors s'y intéresser et ce jour-là, vous remporterez le gros lot.

À propos de Tableau

Tableau Software aide les utilisateurs à visualiser et à comprendre leurs données. Tableau permet à chacun d'analyser, de visualiser et de partager rapidement des informations. Plus de 15 000 comptes clients font confiance à Tableau pour obtenir rapidement des résultats, au bureau ou en déplacement. Tableau Public permet à des dizaines de milliers d'utilisateurs de partager des données sur leurs blogs et sites Web. Pour découvrir comment Tableau peut vous aider, téléchargez la version d'essai gratuite à l'adresse www.tableausoftware.com/trial.





Ressources supplémentaires

Télécharger une version d'évaluation gratuite

Ressources connexes

Création d'une culture dirigée par les données : un rapport spécial établi par The Economist Intelligence Unit et Tableau

Données volumineuses : la prochaine révolution industrielle

eBay utilise l'analyse pour conduire ses affaires

Tableau Software et les Big Data

Aberdeen Group : Maximizing the Value of Analytics and Big Data

Voir tous les documents de présentation technique

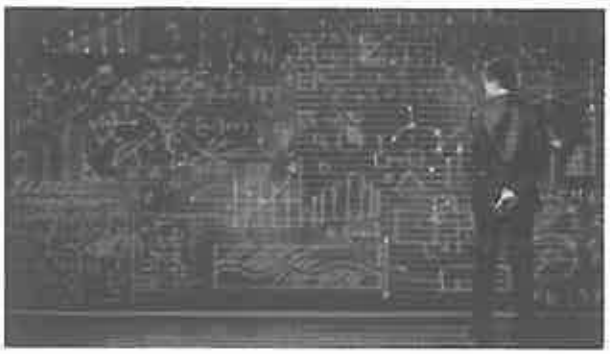
Explorer les autres ressources

- Démo produit
- Formation et tutoriels
- Communauté et assistance
- Témoignages de clients
- Solutions

Document N°12

Le saviez-vous ? Pour le marketing, le Big Data doit devenir Smart Data

Posté le 10 Octobre 2013 par Adobe France



L'ex-Pdg de Google, Eric Schmidt, aime rappeler que si l'on numérisait toutes les communications et les écrits depuis l'aube de l'humanité jusqu'en 2003, il faudrait 5 milliards de gigabits pour les mettre en mémoire. Aujourd'hui ce volume d'informations numériques est produit en deux jours !

Le Big Data pour les marketeurs ?

Au-delà des définitions techniques et traditionnelles, le Big Data est pour les marketeurs une incroyable accumulation de données sur les consommateurs. Dans un monde de plus en plus digitalisé (musiques, textes, conversations, comportements...), le volume des données à traiter continue de croître de façon inédite. **Face à ce constat, le vrai challenge des marketeurs d'aujourd'hui ne réside pas dans la gestion du Big Data mais plutôt dans l'utilisation intelligente de ces données pour découvrir des informations à valeur ajoutée sur chacun des consommateurs.**

Du Big Data au Smart Data ...

Selon l'institut Gartner, le Big Data est à l'origine de 28 milliards de dollars d'investissements dans le monde en 2012 et devrait atteindre 34 milliards de dollars en 2013. Le marketing a une part croissante dans ces investissements et pourtant, les marketeurs expriment une certaine déception face à des volumes et coûts considérables. Ces derniers ont compris l'enjeu : **ce n'est pas le volume de données (Big Data) qui leur permettra de créer de la valeur pour leurs clients, c'est l'intelligence de ces données (Smart Data) !**

Le Smart Data est le processus qui permet de passer des données brutes à des informations ultra qualifiées sur chacun des consommateurs. L'objectif est d'avoir une vision à 360°des clients reposant sur des informations collectées à travers des jeux concours, les réseaux sociaux, les achats lors des passages en caisse, l'utilisation des applications mobiles & la géolocalisation…

Pour y parvenir, les entreprises doivent se doter d'une plate-forme de marketing cross-canal capable de stocker et d'analyser chaque information, pour pousser le bon message au meilleur moment et pour chaque consommateur. L'objectif final est de séduire de nouveaux clients, d'augmenter leur satisfaction et leur fidélité, et permettre ainsi une amélioration considérable du ROMI.

Allianz, un exemple réussi de passage du Big Data au Smart Data

Allianz utilise la plate-forme de marketing conversationnel Neolane pour obtenir une vision à 360° de ses clients grâce à une base de données centralisée. Cibler de manière fine selon plusieurs critères tout en respectant les préférences clients, gérer la personnalisation, la pertinence et la cohérence des messages cross canal délivrés par email, courrier, web et call center sont devenus des priorités.

« Depuis que notre stratégie marketing repose sur la gestion du Big Data, nous sommes passés de larges campagnes d'assurance à des campagnes finement ciblées, très pertinentes grâce à une personnalisation 1to1, basée sur l'utilisation des nouvelles informations découvertes sur les clients. » indique James Horsburgh, Marketing Database Manager chez Allianz Retail, UK.

Allianz a ainsi réussi à instaurer un véritable dialogue avec chacun de ses clients. Un reporting précis permet ensuite à l'assureur de mesurer efficacement les performances marketing et commerciales de la marque. Les résultats sont au rendez-vous : la meilleure qualité des données, la possibilité d'avoir une segmentation comportementale très fine et de déclencher des messages pour chaque client ont permis à Allianz de générer des revenus additionnels dont le total est un nombre à plus de sept chiffres.

Document N°13

Ne manque-t-il pas un V au Big data ?

Écrit par Patrice Poiraud Directeur Big Data IBM
Article paru le 10/02/2014

LE CERCLE. En 2013 le Big data sera-t-il toujours l'un d'un des sujets majeurs au cœur des préoccupations des entreprises ? Il semble que oui si l'on en croit les 10 tendances technologiques stratégiques en 2013 révélées par Gartner. Même si le Big data semble s'imposer dans le paysage numérique il semble bon de préciser quelques points.

La véracité, un élément crucial...

Premièrement, le Big data ne signifie pas uniquement un grand volume de données. Selon la définition initialement mise en avant par l'analyste Gartner, Doug Laney, en 2001, le Big data implique trois dimensions : le volume, la vitesse et la variété des sources. À cette définition, il convient d'ajouter un quatrième "V" : la véracité, soit la confiance en l'information. Mais pourquoi ?

Si les entreprises ont opéré de réelles avancées dans la structuration et l'analyse des données internes, les informations externes qu'elles soient structurées ou non constituent encore un vaste chantier auquel elles ne sont que très peu préparées.

Pourtant le Big data se caractérise aussi par ces données externes et non structurées issues d'internet, des réseaux sociaux, mais également de capteurs ou de logs. Ces données majoritairement générées par les consommateurs permettent de retracer leurs interactions avec une marque grâce à internet, aux réseaux sociaux, blogs et autres plateformes de partage : les clients ont maintenant la parole produisant ainsi de plus en plus d'informations non structurées et aussi utiles à l'entreprise.

Et justement parce qu'elles sont émises directement par les clients, elles nécessitent un travail sur la qualité et l'interprétation. Ces données externes ont donc une dimension comportementale. L'exploitation de ce nouvel "or noir" qu'est le Big data n'est pas sans difficulté et pose de nouveaux défis. En effet, toutes les informations n'ont pas la même valeur et certaines peuvent s'avérer peu voire pas pertinente (ou même erronées).

Raison pour laquelle une majorité des entreprises se méfie de leur véracité et de leur validité comme le corrobore le rapport 2012 The real-world use of Big data (*) par IBM et la Saïd Business School de l'université d'Oxford.

Ce rapport montre que moins de la moitié des entreprises engagées dans une initiative Big data collectent et analysent les sources externes de données. Ce phénomène s'explique par le fait que les dirigeants d'entreprise craignent l'incertitude et le manque de véracité inhérents à certains types de données, telles que la météo, l'économie ou le ressenti et la sincérité des personnes qui s'expriment sur les réseaux sociaux.

Lorsque l'on traite ce type de données, on aura beau procéder à un tri en amont, une part d'incertitude subsistera. Pourtant, malgré cette incertitude, ces données non structurées sont une source d'informations précieuses. La nécessité de reconnaître et d'accepter cette incertitude comme une caractéristique du Big data est indispensable.

Un exemple de cette incertitude s'illustre dans le secteur de la production d'énergie : la météo est incertaine, mais une entreprise de service public doit prévoir ses capacités de production. Dans de nombreux pays, des organismes de réglementation exigent qu'un pourcentage de la production provienne des sources renouvelables, pourtant ni le vent ni les nuages ne peuvent être prévus avec précision. Alors, comment prévoir ?

Pour gérer l'incertitude, les analystes ont besoin de créer un contexte autour des données. Une façon d'y parvenir est de combiner les données avec des sources moins fiables.

Vestas, numéro 1 mondial des éoliennes, constitue un exemple en la matière. Confrontée à la problématique constante d'optimiser l'emplacement des éoliennes pour être au meilleur endroit pour capter les vents les plus forts possible, cette entreprise a opté pour une solution lui permettant d'analyser 16 petaoctets de données pour chaque éolienne ainsi que des données météo concernant plus de 170 paramètres. Vestas a pu réduire la durée de prévision de la vitesse des vents de trois semaines à moins d'une heure.

Et surtout a pu considérablement réduire la surface de prospection. Avant, il lui fallait travailler sur des zones de 27 kilomètres carrés pour trouver l'emplacement idéal. Dorénavant, la surface à considérer n'excède pas 3 kilomètres sur 3 et Vestas est en mesure de positionner idéalement ses installations par rapport aux vents et donc de maximiser leur rendement.

... pour que le Big data soit un accélérateur de business.

Deuxièmement, les entreprises utilisant le Big data et notamment l'analyse des données externes est deux fois plus performantes que celles opérant dans le même secteur et qui ne l'ont pas fait. Il est démontré que le chiffre d'affaires de ces entreprises est en moyenne 1,6 fois plus élevé par rapport à celui des entreprises n'ayant pas intégré de solutions d'analyses. Pourquoi me direz-vous ?

Ces entreprises ont tout simplement la capacité de prendre les bonnes décisions au bon moment... Si le Big data permet d'améliorer l'avantage concurrentiel des entreprises, il est indispensable de passer de l'initiative business à l'impératif business afin d'avoir une meilleure connaissance de son marché et de ses clients. La véracité est donc au cœur de l'analyse du Big data. Une véracité qui, plus généralement, permettra aux entreprises d'être sur un pied d'égalité avec le consommateur.

Le dernier point à souligner sera, finalement, qu'il ne suffit pas de simplement "gérer" les données. Il faut les exploiter activement et de façon à ce qu'elles aient un impact concret sur les performances d'une entreprise. Catalina Marketing a par exemple mis en place, pour ses clients, une stratégie de coupons de réduction ciblés en analysant en temps réel les tickets de caisse. Cette stratégie a conduit à doubler la rentabilité des campagnes de couponning.

L'idée du Big data est de combiner différentes sources pour avoir une vue à 360° et surtout la plus véridique du client.

(*) The real-world use of Big data par IBM et la Saïd Business School de l'université d'Oxford -2012

ÉPREUVE N° 33