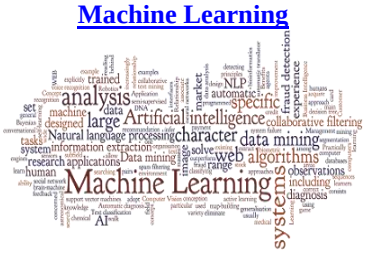
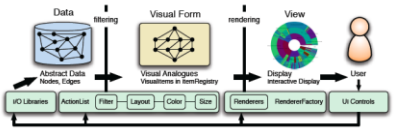
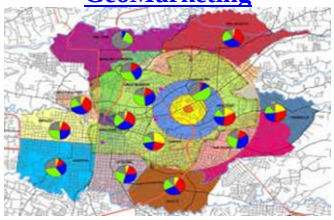
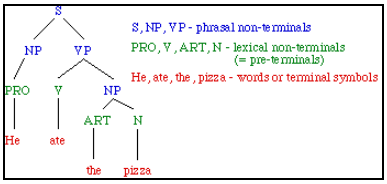
Premiers Pas dans les Big Data

Michel Bruley

Text Mining 	**Sentiment Analysis** 	**Social Network** 
Web log & Clickstream 	**MapReduce** 	**Marketing Attribution** 
Social CRM 	**Churn** 	**Machine Learning** 
Product Affinity 	**Next best offer** 	**Data Visualization** 
GeoMarketing 	**Pricing** 	**Natural Language Processing** 

Premiers Pas dans les Big Data

1 - L'utilisation des big data va-t-elle révolutionner le Marketing ?

11 - Les big data permettent de mieux comprendre les clients

111 - Exploitation analytique des textes

112 - Analyse des opinions et des sentiments

113 - Analyse des réseaux sociaux

114 - Comprendre le parcours du client sur le net avant qu'il achète

115 - Analyse des affinités produits

12 – Le big data permettent d'agir mieux.

121 - Buzz Marketing

122 - Elaborer la prochaine meilleure offre à faire à un client

123 - Répartition des budgets marketing en fonction des comportements clients

2 - Les big data boostent les avantages concurrentiels

21 - eBay est déjà bien équipé pour le big data

22 - Qu'est-ce que big data veut dire chez LinkedIn ?

23 - Big data, les pionniers nous montrent la voie

24 - Une solution big data pour traquer la fraude dans une salle de poker en ligne

25 - Quelques aperçus sur l'expérience big data de Barnes & Noble

26 - Quelques usages de big data expérimentés par SuperValu

27 - Razorfish analyse des big data et crée des expériences clients profitables

28 - De l'expérience big data de Gilt groupe

29 - Les aventures de Wells Fargo dans le big data

3 - Big data : au-delà des thèmes métiers et des premiers cas d'utilisation

31 - Big data : un nouveau champ de travail pour les experts du décisionnel

32 - Infrastructure big data : répondre à des exigences de volume, de variété et de vitesse

33 - De la préparation des big data pour les analyses avancées

34 - Hadoop n'est pas la panacée universelle

35 - Des big data pour mieux servir les clients

36 - Big data et traitement automatique du langage naturel

37 - Big data, commerce électronique et cloud computing

38 - Big data : information, propagande, désinformation & mystification

Annexes

1 - L'utilisation des big data va-t-elle révolutionner le marketing ?

Sans conteste, c'est la fonction Marketing qui s'est lancée le plus tôt et de la façon la plus importante dans l'usage des big data. Elle en avait besoin pour renouveler ses pratiques et faire face à un contexte compliqué, entre autres du fait de la globalisation des activités au niveau mondial et du développement de nouveaux comportements des clients (activités multicanal par exemple), facilité par les nouvelles technologies (internet, mobilité, …). Dans ces conditions la possibilité d'analyser les big data est une opportunité de mieux comprendre les conditions du jeu des affaires, et d'agir de façon plus pertinente (positionnement, promotion, …).

Cependant cela va-t-il révolutionner le Marketing ? Faut-il jeter au feu toutes les approches traditionnelles, marketing mix, 1to1 marketing, etc…. ? Pour moi l'utilisation des big data n'apporte pas d'idée nouvelle, de concept nouveau, mais permet de mieux mettre en œuvre certaines actions, par exemple en améliorant la connaissance du client (comportement web & multicanal, affinité produit, sentiments, réseaux sociaux, etc. …), on peut envisager faire des propositions plus pertinentes tant sur le fond que dans la forme ou le moment (retargeting, cross/up selling, anticipation de l'attrition de la fraude, des risques, tarification dynamique, etc….).

11 - Les big data permettent de mieux comprendre les clients

111 - Exploitation analytique des textes

Les entreprises cherchent de plus en plus à tirer parti des big data, en particulier des données textuelles, celles générées via les outils utilisateurs par les applications bureautiques ou web. Les analystes spécialisés sur le sujet pensent que 70 % des informations qui intéressent les entreprises sont nichées dans les documents word, excel, les courriels, etc. Ces données ne sont pas prédéfinies dans un modèle et ne peuvent pas être parfaitement rangées dans des tables relationnelles. Elles se présentent le plus souvent sous une forme très libre, mais contiennent des dates, des chiffres, des mots clés, des faits qui peuvent être exploités.

Un nouveau défi pour les entreprises en matière d'analyse de données est donc de significativement progresser dans l'exploitation de ce type de données non structurées. En matière de connaissance client par exemple, il s'agit en particulier de mieux exploiter les archives des propositions commerciales et des contrats ou d'écouter les conversations web ou de tirer parti des dialogues via les courriels. La maîtrise des relations, notamment des discussions de l'entreprise avec sa communauté de clients et les acteurs de son écosystème, est une clef du marketing actuel qui est en pleine mutation du fait des nouvelles technologies (mobilité, médias sociaux, ...).

La quantité de ce type de données numériques exploitables est en croissance permanente et comme « l'extraction manuelle » d'informations est extrêmement ardue, voire pratiquement impossible à grande échelle, le recours à des outils informatiques spécifiques pour le traitement de données textuelles non structurées s'impose. C'est ainsi que sont nés, les outils de fouille de données textuelles, qui permettent d'automatiser le traitement de gros volumes de contenus texte, pour répertorier de manière statistique les différents sujets évoqués et en extraire les principales informations.

La fouille textuelle applique sur les textes des traitements linguistiques, notamment morphologiques, syntaxiques, sémantiques, ainsi que diverses techniques d'analyse de données, de statistique, de classification, etc. Concrètement il s'agit de synthétiser (classer, structurer, résumer, …) les textes en analysant les relations, les structures et les règles d'association entre unités textuelles (mots, groupes, phrases, documents). Au final cela permet d'automatiser la production et la gestion de documents (notamment des résumés) ou d'informations (extraction, recherche, diffusion).

La fouille textuelle a de nombreuses applications par exemple dans le domaine de la relation client, elle permet en particulier d'explorer le contenu de

documents (par exemple les questions ouvertes dans une enquête, les commentaires et plaintes des clients, l'analyse des réclamations de garantie) ; affecter des documents à des thèmes prédéfinis (redirection, filtrage des courriels, organisation des documents par catégories, classement des contacts au centre d'appel) ; composer des résumés de textes (abstraction et condensation) ; interroger des textes par concepts, mots-clés, sujets, phrases visant à obtenir des résultats triés par ordre de pertinence, à la Google ; et enfin augmenter la performance de modèles prédictifs en combinant des données textuelles et des données structurées.

Teradata qui a un long passé dans l'analyse des données ne pouvait pas manquer ce passionnant domaine. Comme les solutions classiques sont mal adaptées pour certains traitements nécessaires pour ces informations textuelles, Teradata a acquis au début de 2011 la société Aster Data qui dispose d'une solution spécialisée brevetée SQL-MapReduce ™. Avec ce moyen supplémentaire qui permet de mieux exploiter de grands volumes de données non relationnelles, Teradata est en mesure de proposer à ses clients des solutions d'analyse très innovantes.

Les solutions Teradata Aster peuvent aider les entreprises à traiter les données textuelles brutes, appliquer une variété d'approches analytiques et chercher des informations et de la signification dans les textes, afin par exemple de : surveiller les commentaires des clients à travers de multiples canaux pour comprendre leur perception et leur satisfaction ; identifier les domaines de préoccupation et d'intérêt dans les discussions des clients ; identifier les tendances dans le développement de la fraude ; comprendre et influencer la façon dont les marques et les entreprises sont perçues dans les forums en ligne (blogs, médias sociaux, etc.) ; identifier les tendances dans les plaintes et les retours qui révèlent des types de défaillances ; classifier et indexer des documents pour faciliter la recherche et la récupération ; ou enfin analyser les enregistrements des appels et des plaintes pour identifier les clients rencontrant des problèmes de qualité et qui risquent de partir à la concurrence.

Pour conclure, la fouille textuelle est un ensemble de technologies qui permet de détecter des éléments de langage, de les transformer en un type de données qui peuvent être manipulées et faire l'objet de traitement statistiques. Pour aller plus loin, vous pouvez cliquer sur le lien ci-dessous, pour découvrir pourquoi il est nécessaire d'utiliser des outils analytiques avancés tels que ceux de Teradata Aster, pour exploiter pleinement les données non structurées. Des entreprises des secteurs de la distribution ou du web comme Barnes & Noble ou LinkedIn utilisent déjà des solutions Teradata Aster pour obtenir des avantages concurrentiels.
http://www.asterdata.com/product/faq.php

112 - Analyse des opinions et des sentiments

Les analyses de textes mettent en lumières deux types principaux d'information, des faits et des opinions. La plupart des méthodes actuelles de traitement des informations textuelles ont pour objectifs d'extraire et d'exploiter des informations factuelles, c'est le cas par exemple des recherches que nous faisons sur le web. L'analyse des opinions s'intéresse quant à elle aux sentiments et émotions exprimés dans les textes, elle se développe beaucoup aujourd'hui du fait de la place prise par le web dans notre société, et du très grand volume d'opinions exprimées quotidiennement par les consommateurs grâce à l'avènement du web 2.0.

En quoi consiste l'analyse des opinions ? Il s'agit d'identifier l'orientation d'une opinion exprimée dans un morceau de texte (blog, forum, commentaire, site web, document sur un site de partage, etc.). Autrement dit, il s'agit de déterminer si une phrase ou un document exprime un sentiment positif, négatif ou neutre, concernant un objet défini. Par exemple dire : « Le film était fabuleux », est l'expression d'une opinion, alors que dire « l'acteur principal du film est Jean Dujardin », est la formulation d'une donnée factuelle. L'analyse des opinions peut se faire à différents niveaux. Au niveau du mot : le film est distrayant et motivant ; au niveau de la phrase : la police (sujet) traque (verbe) la contrebande (objet) ; ou enfin au niveau du document, c'est-à-dire d'un ensemble de phrases : ses premiers films étaient très bons, mais celui-là ne vaut rien.

En fait une opinion peut être caractérisée par une

formule de cinq composants, le quintuple : Oj, Fjk, Hi, Tj, SOijkl ; où Oj est un objet cible ; Fjk une caractéristique de l'objet cible ; Hi un porteur d'opinion ; Tl le moment où l'opinion est exprimée et SOjkl est l'orientation de l'opinion, du porteur d'opinion Hi, au sujet de la caractéristique Fjk de l'objet Oj au moment Tl. En utilisant cette formule on peut ainsi structurer un ensemble documents, de données web non structurées, en mettant en lumière tous les quintuples compris dans les textes. Les quintuples sont des données structurées qui peuvent être analysées qualitativement ou quantitativement, et être représentées visuellement avec les moyens classiques des systèmes décisionnels. Toutes sortes d'analyses sont possibles. L'analyse des opinions ne consiste pas uniquement à caractériser l'opinion d'une personne exprimée par des mots et des phrases, mais aussi par exemple à comparer les avis de différentes personnes ou groupes.

La première opération de l'analyse des opinions contenues dans un texte consiste à supprimer les phrases qui ne contiennent que des faits, pour ne retenir que celles qui expriment des opinions et en définir la polarité (positive, négative ou neutre). Concrètement vous avez des adjectifs qui indiquent des faits (rouge, métallique), ou des sentiments positifs (honnête, important, mature, grand, patient), ou négatifs (nocif, hypocrite, inefficace) ou subjectifs sans être ni positifs, ni négatifs (curieux, étrange, bizarre, sans doute, probable). Il en est de même pour les verbes, positifs (louanger, aimer), négatifs (blâmer, critiquer), subjectifs (prédire) ou les noms positifs (le plaisir, la jouissance), négatifs (la douleur, la critique) et subjectifs (la prédiction, l'impression).

Attention, définir le sens d'une suite de mots ou d'une phrase peut parfois être compliqué. Un homme grand ne doit pas être confondu avec un grand homme, et la ponctuation qui a une grande importance, peut jouer des tours : Le cyclope dit, "Ulysse est idiot", n'a pas le même sens que "Le cyclope, dit Ulysse, est idiot". Il faut aussi tenir compte que des mots ou des phrases peuvent signifier des choses différentes en fonction des contextes et des domaines, ou de la subtilité de l'expression des sentiments lorsque quelqu'un fait de l'ironie par exemple.

Au final cependant, l'analyse des opinions et des sentiments est à même d'apporter beaucoup d'informations sur les populations étudiées, et les responsables marketing avertis savent déjà en tirer partis. C'est le cas de nombreux clients de Teradata Aster, comme Barnes & Noble, LinkedIn, eBay par exemple. Pour aller plus loin sur ce sujet vous pouvez utilement consulter le site suivant : http://www.asterdata.com/solutions/data-science.php

113 - Analyse des réseaux sociaux

Un réseau social est une structure sociale qui lie des acteurs entre eux (des individus ou des organisations), et met en lumière la manière dont les acteurs sont connectés, allant de la simple relation aux liens familiaux. Nous participons tous à de nombreux réseaux qui correspondent à des dimensions de notre vie (famille, étude, travail, activités de loisirs). Notre appartenance, nos activités, notre place dans ces réseaux sont pour les marketers une source intéressante d'informations, de connaissances et de possibilités d'actions pour promouvoir leur offre, selon le principe que les comportements des individus sont en partie liés aux structures dans les lesquelles ils s'insèrent.

Internet a favorisé le développement, le fonctionnement de réseaux sociaux et pour les exploiter à leur profit, les marketers se sont appropriées les techniques d'analyse nécessaires. En effet un réseau peut être représenté comme un graphe et être mathématiquement analysé. Dans ces approches, les acteurs sont des nœuds et les relations sont des liens, formant ainsi un modèle où tous les liens significatifs peuvent être analysés via la construction d'une matrice pour représenter le réseau. On peut alors obtenir un graphe à l'aide de traitements mathématiques effectués sur les matrices, et rechercher entre autres, la présence de clique, de chaîne, de cycle pour caractériser le réseau. Enfin à l'aide d'algorithme on peut calculer les degrés de force et de densité entre les entités sociales, pour par exemple déterminer le capital social des acteurs.

Il existe de nombreuse mesures des connexions, des distances, du pouvoir, du prestige : le nombre de

nœuds, le nombre de liens présents vs le nombre liens possibles, la somme des liens vers les autres membres, le degré de densité, le degré de cohésion, l'intermédiarité, la longueur des chemins, le degré avec lequel n'importe quel membre du réseau peut atteindre les autres membres du réseau, les trous structuraux, etc. Ainsi peut-on caractériser à la fois un réseau et chaque acteur, pour par exemple identifier des personnes clés qui ont un rôle important en matière de communication ou d'influence.

L'analyse des réseaux sociaux est précieuse par exemple pour contrôler les flux d'informations, améliorer / stimuler la communication, améliorer la résilience d'un réseau, trouver des communautés, ou pour faire confiance. Pour les marketers c'est une opportunité de mieux connaître, cibler, approcher ses clients, prospects, suspects, pour leur vendre plus, pour mieux animer des communautés, pour innover, pour se différencier de la concurrence et développer un avantage concurrentiel.

Des entreprises comme Myspace, LinkedIn ou Mzinga ont bien compris l'intérêt de ce type d'approche et pratiquent déjà largement l'analyse de réseaux sociaux, pour lancer, tester de nouveaux produits, améliorer les expériences de leurs clients et mieux les satisfaire. Mzinga en particulier dont l'activité consiste à fournir des moyens pour animer des communautés de clients, propose des outils d'analyse des réseaux. Ainsi les 14 000 communautés, regroupant 40 Millions de personnes qui sont gérées avec des outils de Mzinga, peuvent être analysées par leurs animateurs et permettre d'en optimiser le fonctionnement.

Mais attention, pour faire de l'analyse de réseaux sociaux il faut d'autres solutions que celles des approches décisionnelles classiques, fondées sur des bases de données relationnelles et des outils de BI. Les entreprises citées ci-dessus qui pratiquent déjà ce type d'analyse ont dû développer par elles même leur application. Elles ont dû en particulier recourir à de nouvelles solutions type Hadoop et/ou Teradata Aster et mettre en œuvre des programmes MapReduce qui supposent des infrastructures spécifiques et des spécialistes de ce type d'analyse de données. Pour aller plus loin sur les expériences des entreprises citées plus haut, vous pouvez utilement accéder à des informations les concernant en suivant le lien ci-dessous :
http://www.asterdata.com/customers/index.php

114 - Comprendre le parcours du client sur le net avant qu'il achète

Le marketing est en cours de redéfinition par l'évolution des habitudes des consommateurs, les choix presque sans limite pour le placement de publicités et un meilleur accès aux clients à travers une variété de canaux. En conséquence, de nombreuses entreprises modifient leur plan d'actions et la répartition de leurs budgets pour les différents canaux, notamment le web, les campagnes, notamment sur les mobiles, les médias sociaux, etc.

Les études qui ont récemment analysées les parcours des acheteurs en ligne montrent que les actions classiques de marketing sur le web comme l'achat de mots clés auprès de moteurs de recherche, la multiplication des bannières, le recours aux plateformes d'affiliation, les campagnes d'e-mailing restent les solutions privilégiées pour stimuler les ventes en ligne, alors que les investissements réalisés dans les médias sociaux restent très secondaires, utilisés pour faire le buzz et rarement associés à la génération de revenus.

Le problème des e-commerçants est de déterminer les meilleures approches pour attirer des acheteurs web tout au long de l'année, et à certaines périodes clés, comme les fêtes de fin d'année par exemple. Il s'agit pour eux de comprendre ce qui pousse les consommateurs à acheter ? Comment découvrent-ils les offres ? Font-ils des recherches ou non, approfondies ou non ? Quelle place le buzz social a-t-il dans le processus d'achat ?

Les consommateurs qui achètent en ligne, ont forcément été influencés (un peu, beaucoup, ..., pas du tout) par les actions marketing sur le web des e-commerçants. Alors que le trafic organique (le fait que le client accède directement au site) est le canal le plus rentable pour commercer, il est également le moins assuré parce que la plupart des acheteurs ne connaissent pas la plupart du temps l'adresse du site.

Généralement les acheteurs passent par diverses étapes avant d'accéder au site et réaliser une transaction. En fait ils ont dans 80% des cas fait des recherches, réagis à un courriel promotionnel, vus une publicité ou un article, ou utilisés un comparateurs. Alors que la majorité des acheteurs sont touchés par un programme de marketing avant d'effectuer leur achat, 45% sont exposés à au moins deux actions de commercialisation avant la finalisation de leur transaction. Ceci montre l'intérêt qu'ont les commerçants à ne pas se contenter de connaître uniquement la dernière action effectuée par leur client (le fameux dernier clic).

Alors que les commerçants emploient une variété d'outils, et développent un grand nombre d'actions de promotion et de communication, les études montrent que les investissements dans les moteurs de recherche et les e-mailings sont les plus efficaces pour générer du chiffre d'affaires. Suivant les industries ces deux types d'actions génèrent à eux seuls de 40 à 60% des ventes. L'impact des autre actions, affiliation, bannières, ...etc., est plus difficile à évaluer car il est souvent très en amont dans le processus d'influence, et peu d'études ont des données sur une profondeur historique suffisante pour les mettre en lumière. Comme la plupart des commerçants se fondent sur des analyses du dernier clic, l'évaluation des apports des différents types d'actions est totalement faussée.

De plus en plus d'entreprises souhaitent mieux comprendre l'ensemble du parcours de leur client avant l'achat, ainsi que la réelle influence de leurs efforts de marketing. Ce niveau de connaissance exige d'appréhender de façon plus exhaustive les relations que le client a eues avec la marque, et de dépasser les simples statistiques liées aux derniers clics. Il s'agit de mettre en place un programme d'évaluation des multiples actions ayant touchées le client avant qu'il achète, en s'appuyant sur une gestion et une analyse adéquates des Big Data correspondantes.

Pour aller plus loin sur ce sujet d'attribuer aux différentes actions marketing la juste part de leur contribution aux résultats, vous pouvez utilement consulter la présentation suivante qui n'a pas l'ambition d'être exhaustive, mais pourrait vous fournir quelques perspectives.

http://www.decideo.fr/bruley/docs/6_mkg_attribution_v0.ppt

115 - Analyse des affinités produits

Identifier les produits qui sont vendus ensemble et utiliser cette information pour mieux définir certains programmes marketing est une démarche très profitable, elle permet par exemple de mieux fixer les assortiments, les communications et les offres promotionnelles. L'analyse des « affinités produits » est une des dimensions de l'analyse des paniers des consommateurs, elle donne la possibilité d'approfondir la connaissance des habitudes des acheteurs et de compléter les analyses traditionnelles : nombre de paniers, panier moyen, variété des achats, sensibilité aux prix, aux promotions, heures de fréquentation, type de paiement, etc.

L'analyse des « affinités produits » participe à la détection des tendances d'achat, des liens entre des produits ou services, des opportunités de ventes croisées, et l'augmentation du chiffre d'affaires. L'apport de ce type d'analyse est de pouvoir identifier, avec un haut degré de précision, le profil des clients susceptibles d'être les plus intéressés par certains produits ou services spécifiques, ou certaines offres groupées.

Cette approche est fondée sur la théorie que si vous achetez certains produits/services, vous êtes plus (ou moins) susceptibles d'acheter certains autres produits/services. Ce qu'achète un client, est considéré comme un ensemble, et l'analyse de paniers cherche à trouver des relations entre les ensembles achetés par différents consommateurs. Au final des relations sont mises en lumière et se présentent sous forme de règle, par exemple : si {bière & pas de repas} alors {chips}. Ainsi sur internet pour générer des ventes additionnelles, des analyses de paniers sont faites pour mettre au point des suggestions du type: «Les clients qui ont acheté le livre A ont également acheté le livre B». Ayant compris que les clients sont très susceptibles d'acheter le shampooing et le revitalisant ensemble, le détaillant ne met pas les deux articles en promotion

en même temps. La promotion d'un seul est susceptible de stimuler les ventes de l'autre.

Les données historiques des paniers sont utilisées pour améliorer la connaissance des clients, mieux comprendre ce qu'ils sont susceptibles d'acheter et de ne pas acheter, et pour créer des programmes marketing plus efficaces. Les historiques des achats servent à identifier les produits/services acquis ensemble par des types de clients définis. Des analyses prédictives sont réalisées pour découvrir parmi les non clients ceux qui sont les plus susceptibles d'acheter les produits/services ciblés et de répondre favorablement à des campagnes spécifiques de ventes croisées. Des associations de produits sont aussi déterminées pour construire des offres groupées. Les produits ou les associations de produits/services qui ne plaisent pas sont mises en lumière et cette information est utilisée pour ne pas promouvoir des offres non désirées. Au final tout cela permet d'augmenter les revenus et de réduire les coûts en ciblant de façon plus précise, plus économique, les clients les plus susceptibles de répondre favorablement aux campagnes.

Concrètement il s'agit d'établir des liens (associations) entre des enregistrements, ou des ensembles, dans une base de données, de trouver les éléments qui impliquent la présence statistique d'autres éléments. Les affinités entre les éléments sont représentées par des règles d'association, par exemple «Lorsque quelqu'un loue un bien immobilier pendant plus de 2 ans et à plus de 25 ans, dans 40% des cas, il achètera un bien ou dans les trois mois qui suivent l'achat d'un bien immobilier, les nouveaux propriétaires vont acquérir des articles ménagers : cuisinières, congélateurs, machines à laver, etc. ».

Techniquement il convient de faire des analyses sur des séries temporelles d'événements et de découvrir des liens dans des séries séquentielles de transactions. Or faire cette analyse via un programme SQL et une base de données relationnelles classiques, est bien sûr possible mais n'est pas optimum, et implique par exemple l'écriture de programmes longs (longs à écrire, longs à exécuter), générant de multiples lectures de la base, ce dernier point étant très pénalisant lorsque la volumétrie des données est importante. C'est pourquoi aujourd'hui on a recours à des solutions MapReduce type Hadoop ou Aster Data, qui peuvent se permettre de ne lire qu'une seule fois la base pour fournir le résultat de l'analyse, via des programmes beaucoup plus simples à écrire et à maintenir.

Avec les nouveaux moyens cités ci-dessus, les analyses des « affinités produits » connaissent un grand développement actuellement. En effet l'analyse de l'affinité peut être faite à tous les niveaux de la hiérarchie produit (produit, famille, rayon, univers), à l'intérieur d'une famille ou à l'extérieur, elle peut porter sur des attributs articles (produits bio, produits nouveaux, régionaux), ou tenir compte des évènements liés aux produits (promo, mise en avant, changement de prix). L'affinité temporelle peut être inversé entre la semaine et le weekend, la semaine j'achète du low cost et le weekend je me paye des extras, elle peut être déclinée sur des saisons particulière, elle peut être ajustée en fonction du cycle de vie (étudiant, travailleur, chômeur, indépendant, travailleur, retraité, rentier), l'affinité peut se faire entre les canaux : super, hyper, proxi, web, etc.

Pour aller plus loin sur ce sujet vous pouvez suivre le webcast « Teradata Aster Big Analytics Appliance » : http://www.asterdata.com/webcasts/big-analytics-appliance.php

12 – Le big data permettent d'agir mieux.

121 - Buzz marketing

Qu'est-ce que le buzz marketing ? Au sens strict du terme, le buzz marketing (anglicisme venant de « bourdonnement » d'insecte) est la création de bruits autour d'un produit, un service, une entreprise ou une marque. Par exemple vous pouvez recruter des consommateurs, de préférence des proactifs bénévoles qui sont des influenceurs auprès de leurs pairs, à qui vous faites essayer vos produits dans de bonnes conditions, avant de les pousser à parler de leur expérience.

Le buzz est l'une des forces les plus puissantes sur le marché, et savoir maîtriser ce canal marketing important est critique. Le bouche à oreille est plus crédible que le vendeur le plus sincère – il touche plus de gens, plus rapidement que la publicité, le publipostage ou même qu'un site internet. C'est cette crédibilité qui donne au bouche à oreille une partie de son pouvoir. Mais attention il tire aussi sa crédibilité du fait qu'il peut être négatif, et dans ce cas le marketing constate qu'il n'est pas facile à contrôler.

Le buzz est devenu une arme de base dans la trousse du marketing, et elle est utilisée de plus en plus fréquemment. Les meilleurs buzz concernent des produits ou services, dont les consommateurs aiment bien parler. Mettez à disposition un très bon produit, et vos clients heureux vont en parler à leurs amis, collègues et famille, et générer le bouche à oreille que vous cherchez. Buzz ou rumeur : quelle est la différence? La rumeur est une information d'origine inconnue ou cachée qui se propage largement sans être vérifiée. Le buzz parle sans intermédiaire ou publicité. Une rumeur est un «sujet», alors que le buzz est un « moyen ».

Avec le buzz, il s'agit de capter l'attention des clients et des médias, et de faire en sorte que parler de votre marque devient amusant, passionnant et valorisant. Il convient pour cela de savoir lancer et alimenter des conversations. Comme toute campagne publicitaire, la campagne de buzz est fondée sur une idée force. Cette idée doit répondre à un besoin inconscient ou exprimé, elle doit être attrayante et originale pour provoquer l'attention, déclencher un besoin ou un plaisir. Pour lancer un buzz il faut suivre les étapes suivantes. 1° Identifier les personnes clés (influentes dans leurs communautés) susceptibles d'être les vecteurs du message. 2° Valoriser ces personnes par le biais d'une expérience personnelle qui flatte leur ego de manière à les rendre impatientes de diffuser le message. 3° Encourager la diffusion du message en fournissant aux vecteurs, des informations et des moyens pour alimenter le buzz. Les vecteurs peuvent être un réseau, un groupe de personnes liées ou entretenues.

Dans ce type d'approches il convient d'identifier et d'instrumentaliser différents types d'acteurs. Les innovateurs, ces personnes ont l'ouverture d'esprit pour accepter de nouvelles idées loin des aspects traditionnels et de la mode. Le marketing ne doit pas leur donner beaucoup d'attention parce qu'ils sont une minorité. Les « adopteurs précoces », ils sont toujours à la recherche de nouveauté, ils sont attirés par les risques. Ils adoptent ou créent de la nouveauté et la transmettent aux abeilles (majorité précoce). Mais attention, ils sont attirés par des droits exclusifs, des offres spécifiques, ils aiment se sentir parmi les privilégiés. Les abeilles. Elles sont au cœur du buzz, ce sont elles qui peuvent alimenter à grande échelle la chaîne d'informations à l'intérieur de la communauté ciblée. Il faut amener les abeilles à parler de leurs expériences, de leurs découvertes et à partager avec d'autres. Ceci en particulier grâce aux « connecteurs » qui ont un carnet d'adresses complet ou aux « mavens » qui sont des experts dans le domaine considéré, et sont des leaders d'opinion. Enfin il y a aussi le grand public, qui lorsqu'il est touché peut générer un effet boule de neige et les retardataires dont il faut se désintéresser car ils sont attachés à des choses traditionnelles et ne sont pas ouverts aux nouveautés.

Il convient de distinguer le buzz traditionnel et le buzz digital. Les techniques de buzz traditionnel sont par exemple le placement de produits, la diffusion d'échantillon, l'animation d'événements de découverte, notamment dans la rue et le recours au parrainage. L'élément clé dans le buzz traditionnel, c'est le contact et la relation entre les vecteurs et les produits, de sorte que le vecteur peut observer, rester en contact avec les personnes ciblées. Ce que le

vecteur doit faire, c'est parler du produit, de l'idée, du service : c'est buzzer. Le buzz digital, aussi appelé marketing viral est une technique utilisant internet qui permet la diffusion très rapide d'idées, de nouvelles et d'informations sur les produits. Deux moyens par exemple peuvent être utilisés pour diffuser une idée : il s'agit de lancer un message drôle ou étonnant qui circulera rapidement entre les acteurs de l'internet, ou de proposer aux acteurs d'internet via une bonne accroche de visiter un site où ils sont invités à s'inscrire où ils sont recrutés pour faire partie d'une campagne d'information/action.

Pour conduire des opérations de buzz il convient au minimum de disposer d'outils opérationnels (base documentaire, base de données, gestion de campagne, d'enquête, ...) et décisionnel (analyse, segmentation, reporting sur les actions,...). Si des opérations sont développées sur le web, notamment via les médias sociaux, des moyens big data (analyse des parcours sur le web, des réseaux, des textes, des sentiments, ...) peuvent se révéler très performants pour comprendre le marché, identifier les différents acteurs et pour conduire les actions.

Pour aller plus loin sur les moyens et les références big data de Teradata Aster dans le domaine du ciblage des influenceurs, vous pouvez utilement consulter le lien suivant :
http://www.asterdata.com/solutions/social-network-analysis.php

122 - Elaborer la prochaine meilleure offre à faire à un client

Quand une entreprise se pose la question de définir qu'elle pourrait-être la prochaine meilleure offre qu'elle puisse faire à un client, elle est tiraillée entre différentes logiques. Logique interne de chiffres d'affaires, de marge, de promotion du mois, ou logique client d'historique d'achats ou de profil de besoins. Quel que soit le canal d'interaction les entreprises ont besoin d'anticiper ce qu'elles vont proposer. Les forces de vente, les centres d'appels ont besoin d'informations sur les clients et de directives (ventes croisées, etc …), les sites web de règles pour leur moteur de placement de publicité, de recommandations de produits services, de tarification personnalisée.

Les premiers moteurs n'étaient pas à même d'intégrer suffisamment d'information pour faire des offres vraiment personnalisées tenant compte des caractéristiques individuelles d'un client. Ils se contentaient d'exploiter quelques informations disponibles pour pousser une offre possible. Les nouvelles générations de moteurs de recommandation sont à même d'intégrer toutes les informations des canaux de commercialisation, qu'elles concernent les historiques d'achat, les communications entrantes ou sortantes, les parcours sur le web, les visites de sites (fructueuses ou non), et de produire des scores qui vont pouvoir être mixés par exemple avec des données temps réel, lorsque le client est en ligne, pour faire la meilleure offre possible.

Tout ceci est rendu possible grâce aux big data qui permettent des calculs massivement parallèles sur des fermes de serveurs low cost, la mise en œuvre d'algorithmes complexes (graph analysis, text mining, path analysis, etc...), le traitement de données sur des profondeurs d'historiques, des séries temporelles très longues. Il s'agit aussi de mettre en place des profils et d'activer des campagnes dynamiques, grâce à des analyses multidimensionnelles, permettant d'aligner les besoins des clients et les intérêts de l'entreprise avec une mise en priorité et une personnalisation des offres. Il y a de gros profits en ligne de mire, pas simplement de l'optimisation des modèles de data mining déjà en place (attrition, ventes croisées, …), c'est stratégique en particulier dès que l'on touche à la tarification.

Ce qui est particulièrement intéressant c'est que sans que ces approches soient à court terme pour toutes les entreprises, elles concernent cependant toutes les grandes compagnies et toutes les entreprises qui vivent directement du web. On trouve donc dans les références big data de Teradata de grandes compagnies traditionnelles de diverses industries (banque, télécommunications, distribution, manufactures, gouvernement, etc.) avec des entreprises comme par exemple Wells Fargo, AT&T, Sears, General Motors, United Air Force, et des entreprises d'e-buiness de différentes tailles comme

eBay, Overstock, Intuit, Razorfish, Full Tilt Poker, Gilt groupe, etc.

Sans parler de la rareté des ressources humaines nécessaires, l'inconvénient des approches Big Data actuelles est qu'il faut avoir recours à de nombreuses technologies pour couvrir ses besoins. Les entreprises pionnières s'en plaignent, surtout les plus petites, telle que LinkedIn qui a recours à une douzaine de solutions différentes dont Teradata pour couvrir ses besoins de stockage et d'analyse (online, near-line, offline). Les écosystèmes ainsi créés sont difficiles à maintenir compte tenu des évolutions incessantes des éléments qui les composent, et qui sont pour la plupart peu matures. Hadoop par exemple fait l'objet d'une douzaine de projets de développement parallèles, pour faire évoluer les possibilités en matière de stockage, d'analyse, de liens avec des solutions existantes du marché plus ou moins incontournables, notamment en matière de restitution, de visualisation.

Dans ce contexte les grands du monde du décisionnel cherchent tous à proposer des moyens d'intégration des données et des solutions décisionnelles et opérationnelles, car au-delà de l'analyse, ce qui compte c'est l'action, c'est par exemple de pouvoir au final faire l'offre la plus susceptible d'être acceptée par un client particulier. Dans ce domaine Tearadata propose UDA (Unified Data Architecture) qui permet d'intégrer ses solutions décisionnelles (Teradata, Aster), ses solutions Marketing (Aprimo, eCircle) et les solutions de ses partenaires spécialisés (par exemple pour le text mining les solutions d'Attensity) et Hadoop (Hotonworks, …). Pour aller plus loin sur ce sujet vous pouvez utilement consulter le lien suivant :
http://www.teradata.com/News-Releases/2012/Teradata-Integrates-Big-Data-Analytic-Architecture/

123 - Répartition des budgets marketing en fonction des comportements clients

Alors que les responsables marketing voient se multiplier les canaux de communication avec leurs clients, leurs budgets n'augmentent pas, mieux des économies leurs sont demandées. Il n'est donc pas question de faire plus de marketing mais de faire mieux, de trouver le bon ensemble d'actions coordonnées qui apportent le taux de conversion le plus élevé, les profits maximum et satisfait les clients. Pour cela les responsables peuvent se fonder sur l'analyse des comportements des consommateurs, mettre en lumière leurs parcours avant les achats et évaluer l'efficacité de leurs dispositifs de commercialisation.

En matière d'analyse de l'efficacité des actions, les entreprise ont généralement aujourd'hui des pratiques assez simplistes d'évaluation action par action sur des métriques diverses, et en matière d'attribution de la vente, on valorise dans le parcours client, soit le premier contact soit le dernier, qui peut être un appel téléphonique, une visite en magasin ou un publipostage par exemple. Cette méthode simple et facile, ignore en fait le détail du parcours du client, passant à côté d'information clés concernant les interactions des clients avec les canaux, les écrans, les messages dans lesquels l'entreprise a beaucoup investi.

Pour mieux comprendre l'intérêt et l'efficacité des différentes actions développées par le marketing, il faut prendre en compte l'ensemble du parcours du client et analyser l'influence des différentes actions marketing, leur contribution à l'acte final d'achat. Pour les entreprises de commerce en ligne il s'agit d'analyser toutes les relations avec leurs clients grâce aux traces numériques qu'elles génèrent, que l'on peut collecter et analyser. Pour les entreprises qui ont aussi des magasins, des centres d'appels, etc. l'approche est la même mais se complique un peu, nécessitant des processus bien définis pour capter toutes les étapes des relations clients.

La prise en compte de l'ensemble du parcours des clients a été hors de portée des responsables marketing jusqu'à très récemment en raison du coût et de la complexité des analyses. Mais grâce aux nouvelles technologies facilitant le traitement des big data de nombreuses entreprises se différencient aujourd'hui de leurs concurrents, grâce à une meilleure connaissance de leurs clients, de leurs parcours et une optimisation de leurs investissements marketing. Pour cela elles cherchent à évaluer l'influence de chacune de leurs actions marketing sur

les comportements de leurs clients, en attribuant à chacune une fraction des achats en fonction de leur place dans le parcours du client.

L'identification des parcours des clients n'est pas triviale et nécessite généralement de traiter un gros volume d'événements temporels. Ceci étant fait, des analyses peuvent être réalisées pour déterminer le meilleur modèle d'attribution des résultats commerciaux aux différentes actions marketing. Les modèles peuvent aller de la simple pondération uniforme de chaque étape d'un parcours, à la mise en œuvre de modèle personnalisé très sophistiqué, en passant par l'attribution d'une valeur grandissante (exponentielle) des étapes successives au fur et à mesure que l'on s'approche de l'étape finale de conversion, ou l'attribution d'un poids particulier à chaque type d'interaction en fonction de ce que les responsables savent déjà de leur modèle d'affaires.

Ces valorisations, même les plus simples, peuvent améliorer de manière significative l'appréciation des contributions aux résultats des différentes actions marketing. Les responsables avisés savent que chaque interaction client peut avoir une incidence sur les autres interactions. Par exemple, les bannières publicitaires et les courriers électroniques ont un impact indirect sur l'utilisation des moteurs de recherche, alors que la recherche via le mobile est souvent liée à l'urgence des achats en magasin. Une évaluation fine de la contribution des actions permet de remettre en cause les dépenses non productives, de mesurer l'impact des interactions sociales, d'optimiser l'impact des campagnes multicanaux, et donc globalement d'orienter le budget marketing pour avoir un meilleur retour sur investissement (ROI).

Pour aller plus loin sur le sujet vous pouvez utilement lire le livre blanc intitulé - L'attribution en Marketing Digital : http://fr.slideshare.net/AT-Internet/lattribution-en-marketing-digital

2 - Les big data boostent les avantages concurrentiels

Les big data boostent la société parce qu'elles sont une des dimensions du grand chambardement provoqué par la numérisation de la société, numérisation de la photo (penser à Kodak), numérisation des textes, des livres, de la presse, de la musique, du cinéma, des télécommunications (Skype), des mesures, etc.. La numérisation produit des données que l'on peut partager plus facilement (google, open data, clients mieux informés, printemps arabe, …), et analyser (Prism/NSA, retargetting, …).

Pour les pionniers les big data sont au cœur même de leur business model. Pour ces entreprises, les big data n'apportent pas un avantage concurrentiel de plus, c'est vital : Cf. Google, eBay, LinkedIn, Critéo, … Cependant, toutes les industries sont intéressées par les big data et en particulier par analyser des données qu'elles ne prenaient pas en compte antérieurement ou à faire de nouveaux types d'analyse. Les big data vont non seulement trouver de nouveaux éléments de réponse à des questions que l'on se pose déjà, mais aussi permettre de formuler de nouvelles questions.

Jusqu'à récemment on savait très bien traiter les données numériques dites structurées (les données des SI des entreprises : facturation, paye, etc.), mais avec cependant quelques limites lorsqu'on avait des processus d'analyse itératif ou des séries temporelles longues. Avec les technologies big data (entre autre Hadoop), on peut travailler toutes sortes de données structurées (faire des itérations, séries longues) ou non structurées (photo) ou à la structure complexe (texte). Par exemple, on traque la fraude différemment si les analyses prennent 90'' au lieu de 90' (Cas de Full Tilt Poker). On peut mettre en œuvre des algorithmes complexes, sur des plateformes low cost (Barnes Noble a ramené à 20' un traitement qui précédemment durait 6h).

Enfin qui va apporter beaucoup, c'est ce que l'on appelle l'internet des objets, tous les capteurs que l'on met partout, tous les objets connectés qui sont en train d'être lancés, avec notamment la géolocalisation, le marketing mobile, … Nous entrons dans un monde où l'on pourra tout mesurer. Le jogger du weekend sort déjà équiper avec sa montre de course à pieds, avec l'open data nous avons accès à toutes sortes de données (les données routières, du cadastre, …) qui viennent enrichir les analyses et nous permettent de mieux comprendre toutes les dimensions de nos activités.

21 - eBay est déjà bien équipé pour le big data

Fondée en Septembre 1995, eBay est un site d'enchères en ligne où n'importe qui peut commercer. Présents dans 190 pays, utilisant 24 devises, eBay compte près de 100 millions d'utilisateurs actifs à travers le monde et près de 300 millions d'inscrits. La société emploie 17 700 personnes et a réalisé en 2010, 9,5 milliards de dollars de chiffre d'affaires net, pour un total de 60 milliards de dollars de marchandises échangées, soit 115K$ de transaction à la minute. Les membres d'eBay du monde entier ont laissé plus de 6 milliards de commentaires d'évaluation au sujet de leurs transactions sur le site, qui enregistre par jour 2 milliards de pages vues, gère 250 millions de requêtes de recherche et 75 milliards d'accès à la base de données.

Dans ces conditions on comprend facilement que le terme big data a un sens chez eBay, qui pour s'éclairer sur toutes ses activités et en tirer un enseignement, utilise un mix de moyens décisionnels fondés sur Teradata et Hadoop. eBay cherche en premier à donner à ses analystes et ingénieurs les outils qu'ils veulent. Les analystes financiers par exemple sont habitués à des outils conviviaux qui ne les obligent pas à programmer et leur masquent les ordres SQL. D'un autre côté beaucoup d'ingénieurs ne sont pas opposés à utiliser le framework de développement MapReduce, qui leur permet de traiter les données non structurées (web logs, text, social network, ...). Enfin l'utilisation parallèle et conjointe de ces moyens crée un environnement analytique particulièrement riche pour les « data scientists ».

La grande quête d'informations vise principalement à comprendre ce dont les clients ont besoin, ce qui fonctionne, ce qu'il faudrait améliorer. Concrètement

des analyses sont faites dans l'optique d'optimiser les expériences des vendeurs et des acheteurs. Par exemple Hadoop s'est révélé particulièrement utile pour interpréter les mots mal orthographiés, ce qui fait que maintenant le moteur de recherche d'eBay sait faire des propositions pertinentes, même si un mot, un nom de produit ont été tapés de façon incorrecte. Toutes les dimensions des relations commerciales sont ainsi passées au peigne fin, promotion marketing, sécurité, service, finance, fidélité, ...dans une recherche constante de qualité. Cependant même si Hadoop offre de nombreux avantages, c'est une technologie difficile à bien maîtriser, et les ingénieurs ont dû retrousser leurs manches et se plonger dans le code source pour en devenir des experts. Au final les résultats sont fondamentalement le fruit d'une collaboration entre les différentes équipes informatiques, d'analyse avancée et métier.

eBay exploite actuellement trois systèmes qui sont alimentés (ELT) grâce à des solutions d'AB Initio et d'UC4. Le premier et le plus petit système, est un entrepôt d'entreprise Teradata de 6 P0 qui intègre des données structurées et peut supporter plus de 500 utilisateurs concurrents. Le deuxième est un « extreme data appliance Teradata » de 40 P0 qui gère des données semi-structurées, permet de réaliser des analyses complexes (saisonnalité, ...) et supporte 150 utilisateurs concurrents. Le troisième est un système Hadoop fondé sur des serveurs de commodité, qui gère plus de 20 P0 de données non-structurées et supporte moins de 10 utilisateurs concurrents. Enfin toutes ces données sont exploitées avec divers moyens dont SQL, Pig, Hive, SAS, Microstrategy, Tableau Software, ...

Pour aller plus loin sur ce cas vous pouvez voir différentes présentations sur Youtube que vous trouverez en tapant les mots : Big Data et eBay

22 - Qu'est-ce que big data veut dire chez LinkedIn ?

LinkedIn qui a été créée en 2003, réalise actuellement 243 Millions de chiffres d'affaires et emploie 1797 personnes. Ce n'est pas ce que l'on appelle une grande entreprise. Cependant LinkedIn a 175 millions de membres répartis dans 200 pays dont 50% en dehors des Etats Unis, deux nouveaux membres se joignent au réseau chaque seconde, et il se dit que tous les « exécutives » des 500 premières entreprises mondiales sont membres. Dans ces conditions, LinkedIn est confrontée à une forte volumétrie de données à traiter. En effet leur système d'information doit supporter par an 2 milliards de recherches effectués par les membres, traiter par jour 75 To de données et 10 milliards de lignes.

En analysant toutes ses données LinkedIn est capable par exemple d'établir le palmarès des mots les plus utilisés par ses membres pour décrire leurs capacités, et ces mots varient d'un pays à l'autre. Aux Etats-Unis et au Canada on met en avant l'étendu de l'expérience, alors qu'en Italie, en France ou en Allemagne on se dit innovant, qu'au Brésil et en Espagne on est dynamique et qu'en Grande Bretagne on met en avant sa motivation.

linkedIn est très certainement une des sociétés qui participent au développement de ce que l'on appelle aujourd'hui dans le monde des affaires la « Science des Données », cette dernière se fonde sur des savoirs faire issus de l'informatique, des mathématiques, de l'analyse de données et du management des affaires. Concrètement il s'agit de pouvoir rapidement collecter des données brutes, les explorer et les analyser, de traduire ces données en informations décisionnelles, et donc globalement de réduire le temps entre la découverte de faits pertinents, la caractérisation d'opportunité métier et le déclenchement d'actions.

Mais qu'est-ce que LinkedIn fait avec ses données ? Elle classiquement fait des analyses pour mieux comprendre et conduire ses activités, mais surtout elle crée des produits / services fondés sur les informations qu'elle génère, soit globalement comme avec les mots les plus utilisés vus ci-dessus, soit individuellement avec des systèmes de recommandations (les gens que vous connaissez peut-être, les emplois qui ...). Les données permettent par exemple : d'identifier des influenceurs et des tendances sociales en matière de viralité ; de tester de nouveaux produits / services, de nouveaux sites pour maximiser l'impact sur l'activité de connexion et l'utilisation du site par les membres ; de comprendre

l'utilisation des services dans le temps en fonction des niveaux d'abonnement, du moyen de connexion (PC, mobile, ...) ; de fournir des rapports détaillés d'analyse des revenus publicitaires ; d'évaluer l'impact d'action de marketing viral ; d'optimiser les moteur de recommandations ; de créer des fonctions spécialisées pour les services pour les entreprises (marketing, recrutement, ...).

Pour pouvoir obtenir ces résultats intéressants de l'exploitation de ses données, LinkedIn a dû développer ses propres applications de gestion des flux de données, de stockage, de recherche, d'analyse de réseaux, etc. et bien entendu ses propres tableaux de bord. Pour cela la société est allée chercher sur le marché les outils ou les solutions dont elles avaient besoin, et l'on peut donc lister de façon non exhaustive : Teradata Aster, Hadoop, Azkaban, Kafka, Project Voldemort, Pig, Pithon, Prefuse, Microstrategy, Tableau software.

Pour aller plus loin à propos du cas LinkedIn, vous pouvez utilement suivre la présentation vidéo de 50' ci-dessous, intitulée « Data Science @ LinkedIn : Insight & Innovation at Scale », de Manu Sharma, Principal Research Scientist and Group Manager, Product Analytics, chez LinkedIn.
http://www.youtube.com/watch?v=W7ZcUJEHAOk

23 - Big data, les pionniers nous montrent la voie

Il y a peu de temps, environ 3 ou 4 ans, si vous vouliez traiter une grande quantité de données textuelles ou de web logs, vous deviez mobiliser de gros serveurs et mettre en œuvre des programmes SQL conséquents, c'est à dire longs à développer et longs à donner leur résultats. Heureusement les demandes étaient peu nombreuses et généralement les volumétries envisagées se mesuraient au plus en téraoctets. Depuis l'e-commerce et les médias sociaux se sont développés, et de nombreuses entreprises voient leurs relations clients et donc leur survie, totalement dépendre de la capacité de leurs moyens informatiques à analyser des web logs et des données textuelles. De plus pour nombre d'entre elles, la volumétrie se compte désormais en centaines de téraoctets voire en pétaoctets comme eBay.

La plupart des jeunes entreprises du monde du e-commerce ou des médias sociaux n'avaient pas les ressources pour mettre en place les solutions évoquées ci-dessus, dont elles avaient besoin. Leurs experts ont donc cherché d'autres voies et développé de nouvelles solutions plus performantes et moins onéreuses, fondées par exemple sur des systèmes de fichiers distribués (DFS) et des programmes MapReduce. Dans ce contexte la solution open source Hadoop implémentée en Java a eu un grand succès, mais il existe aussi d'autres solutions qui permettent de faire du MapReduce. C'est le cas de la solution nCluster de Teradata Aster qui permet de développer des programmes SQL embarquant des algorithmes MapReduce complexes. Ainsi aujourd'hui les entreprises qui veulent traiter des grands volumes de données textuelles ou de web logs complètent à moindre coût leur système d'information décisionnel avec une plateforme analytique spécialisée.

Certains prédisent la disparition des entrepôts de données d'entreprise tels que nous les connaissons aujourd'hui, d'autant plus que des fournisseurs offrent des solutions cloud. Cela ne sera sans aucun doute pas le cas même à moyen terme, et nous verrons donc les entreprises gérer en parallèle différents systèmes spécialisés internes ou externes. Par contre c'est effectivement la fin de l'entrepôt de données centralisé unique qui gère toutes les données de l'entreprise, que d'ailleurs très peu de sociétés avaient réellement mis en œuvre.

En fait les pionniers nous montrent certainement la voie du futur qui est de faire cohabiter les solutions, les nouvelles pour traiter les données multi-structurées et les traditionnelles pour les données structurées, le tout en mode privé ou en mode cloud public. En effet la majorité des solutions sont maintenant fournies sous trois formes: logiciel uniquement, appliance ou cloud, et les pionniers optent pour des solutions hybrides. Le choix entre ces possibilités doit se faire en fonction des exigences spécifiques à chaque entreprise : exigences de la réglementation, de l'industrie, du métier, des relations avec les clients (vie privée), des compétences disponibles, de la sécurité, de l'impact de la localisation des données, etc.

Une des grandes difficultés à court terme que rencontrent les pionniers vient du manque de compétences en matière de big data. En effet l'exploitation de ces dernières relève de ce que l'on appelle la Science des Données, une discipline qui allie les mathématiques, la programmation et le sens des affaires. Pour tirer parti des Big Data il convient donc d'investir dans une équipe ayant ce type de compétences, et de la faire travailler étroitement avec les équipes métiers et informatiques. En effet il est possible de trouver des tendances, des modèles, des segments etc. que l'on ignorait, mais en soi cela ne change rien, il faut transformer ces éléments en opportunités métier et au final en actions concrètes sur le marché. Les experts de la Science des Données savent ouvrir la voie mais ne peuvent pas la parcourir seuls jusqu'au bout.

Parmi les pionniers clients de Teradata on trouve des entreprises de taille très différentes allant de grands groupes genre Wall-Mart, Wells Fargo, Boeing, Apple, avec beaucoup de sociétés liées au web comme eBay, Amazon, Barnes & Nobles, et beaucoup de sociétés beaucoup plus petites comme LinkedIn (1700 personnes), Mzinga (<500), etc. Pour aller plus loin au sujet de Teradata Aster, vous pouvez utilement consulter le lien suivant : http://www.asterdata.com/customers/index.php

24 - Une solution big data pour traquer la fraude dans une salle de poker en ligne

Full Tilt Poker est la seconde plus importante salle de poker en ligne, elle appartient au groupe Rational FT, une société de services informatiques qui développent des logiciels, a des activités de maintenance informatique et réalisent des opérations de marketing digital. http://www.rfts.com/

Un des points fondamentaux de la gestion d'une salle de poker en ligne concerne l'intégrité du site et la confiance des clients. Les joueurs doivent être totalement sûrs qu'ils sont protégés contre la fraude. Toute publicité négative sur ce sujet a un impact direct et immédiat sur la fidélisation de la clientèle, la capacité d'attirer de nouveaux clients, et le chiffre d'affaires. La difficulté ici vient du volume des données à traiter, de la complexité des analyses, de la difficulté de détecter toutes formes de fraude et des coûts afférents. En fonction de l'activité, l'augmentation rapide des volumes de données et les exigences analytiques peuvent impliquer des coûts d'infrastructure accrus et une baisse de la marge globale.

Des fraudes ont lieu dans les salles de Poker en ligne et ne sont pas détectées du fait des faibles capacités d'analyse. L'ampleur même du problème est le plus souvent inconnue. Il est difficile, voire impossible d'identifier de nombreux types de fraude : la connivence entre joueurs, la pratique des multi-comptes, le contournement des systèmes de prévention des déconnections intempestives, l'utilisation de logiciels d'assistance, le changement de joueur physique donc de style de jeu, l'utilisation d'un cheval Troie pour voir les cartes cachées des adversaires, etc. …

La détection des fraudes nécessitent de traiter beaucoup de données et prend du temps, il faut travailler par itération de question en réponse avec des cycles très longs pour aboutir. Au final il est difficile de traquer les fraudes sophistiquées et de nombreuses actions frauduleuses ne sont pas repérées. Cependant il est possible d'identifier les joueurs qui jouent souvent ensemble et sont susceptibles d'être de connivence, qui ont déjà fait des déconnexions intempestives, qui ont joué plus de 500 mains, etc.

En utilisant une plateforme analytique Aster Data, Full Tilt Poker a pu améliorer considérablement sa traque de la fraude, en particulier en pouvant prendre en compte dans les analyses le détail des mains jouées. Précédemment le traitement de ces données étaient long, transfert d'un serveur à un autre, décompression via des traitements par lot et au final l'analyse de sept jours de fraude prenait une semaine au rythme de 1200 mains par seconde. Avec la solution Aster Data, les transferts sont supprimés, les traitements sont massivement parallélisés sur un ensemble de serveurs low cost et l'analyse de sept jours de données prend 15 minutes au rythme de 140 000 mains par seconde.

De plus la plateforme analytique Aster Data est

capable d'exécuter des algorithmes sophistiqués et de détecter des actions frauduleuses dans de très grands ensembles d'événements. Full Tilt Poker a pu observer que des requêtes qui prenaient 90' prennent maintenant 90''. Des requêtes qui n'aboutissaient pas, s'exécutent généralement en moins d'1 heure. Ces capacités ont permis de donner accès aux données, via des requêtes SQL, à tous les analystes de la société.

En résumé, la solution big data d'Aster a permis d'identifier des formes de fraude jusque-là non détectées, de réduire considérablement les temps des cycles d'analyse, de faire progresser extraordinairement les performances en particulier pour le traitement des mains, d'autoriser des analyses plus profondes des fraudes sophistiquées, de mieux répondre aux exigences de traque du blanchiment d'argent, de fournir aux joueurs de nouveaux services de personnalisation des prestations, d'améliorer le reporting interne concernant l'activité de la salle de Poker, tout en réduisant le TCO du système (matériel, maintenance, SAN, personnel, réseau).

Pour aller plus loin sur les capacités analytiques de la solution Teradata Aster vous pouvez utilement consulter cette page :
http://www.asterdata.com/solutions/big-data-analytics-discovery.php

25 - Quelques aperçus sur l'expérience big data de Barnes & Noble

Barnes & Noble fait partie des 500 plus grosses sociétés au monde. Elle gère 1350 librairies (730 magasins de ville et 630 dans des campus), ainsi que la plus grande librairie en ligne sur internet, elle a globalement 10 millions de clients, vend 300 millions de livres par an, et propose un catalogue de 6 millions de références.

Les objectifs de l'entreprise sont de faire mieux que la concurrence (en particulier mieux qu'Amazon) et de dominer le marché de l'eBook. Pour cela l'entreprise vise de bien analyser et maîtriser ses activités à travers ses différents canaux de vente et de développer une approche personnalisée des clients grâce à des segmentations plus élaborées, des programmes marketing adaptés, une meilleure coordination des expériences des clients à travers les canaux, avec des sites web plus réactifs, des moteurs de recommandations en ligne plus optimisés, et un service en magasin plus pertinent.

Barnes & Noble a dû relever plusieurs défis techniques, comme consolider 9 silos de données indépendants, intégrer les données des différents canaux, gérer des données multi-structurées, faire face à l'accroissement de la volumétrie (en particulier celle liée au web), et aussi développer des capacités d'analyse au-delà d'une BI centrée sur le reporting. Pour cela la société a choisi la solution Teradata Aster nCluster pour gérer sa plateforme analytique, pour supporter son nouveau moteur de recommandation, pour traiter ses web logs (20 fois plus rapidement), pour analyser toutes ses données (Coremetrics, SAP, ventes en librairie, vente en ligne, plus 13 autres sources) beaucoup plus facilement (le modèle de churn tourne en 20' au lieu de 5 heures), pour améliorer considérablement la productivité des analystes et enfin pour traquer comment les clients utilisent leur Nook (la liseuse commercialisée par Barnes & Noble aux États-Unis depuis 2009).

Le nouveau moteur de recommandation par exemple permet un scoring dynamique des clients et des affinités produits. Pour cela concrètement il s'agit d'identifier tous les visiteurs (les fidèles et les nouveaux clients), de tirer parti d'une analyse de 186 millions de paires de produits, de tenir à jour des modèles d'évaluation ou des segmentations en fonction des changements des client, de divers facteurs d'environnement, et de mettre en œuvre ces modèles pour toutes les communications clients quelques soient les canaux ou les événements (initiatives des clients ou de Barnes & Noble).

La solution Aster a aussi permis d'améliorer le traitement des données de navigation des visiteurs du site web, notamment en matière d'identification des sessions, d'analyse textuelles et d'étude des parcours. Les marketers sont maintenant mieux armés pour gérer leurs problématiques (panier abandonné, tarification, optimisation des contenus, personnalisation des bannières, …). Par exemple le responsable de la gamme de produits liée à la liseuse Nook, limitait ses analyses en matière d'attrition car

il fallait 6 heures avec l'ancien système pour traiter les 240 millions de lignes et toutes les requêtes imbriquées nécessaires. Maintenant que cette même analyse se fait en 20 minutes, il ne se freine plus.

Marc Parrish, Vice President Customer Loyalty and Retention de Barnes & Noble, résume bien la situation en disant « Quand je suis arrivé à B & N nous ne permettrions pas à beaucoup de personnes d'accéder librement aux données parce que le système de l'époque n'aurait pas eu la capacité à supporter la charge. Maintenant c'est totalement changé. Nous avons démocratisé les données ».

Pour aller plus loin sur ce cas, vous pouvez utilement écouter l'interview filmée de Marc Parrish, via le lien suivant (vidéo 4') : http://www.asterdata.com/barnes-and-noble-video.php

26 - Quelques usages de big data expérimentés par SuperValu

SuperValu Inc est une société de distribution américaine. La société, dont le siège est à Eden Prairie dans le Minnesota, existe depuis plus d'un siècle. Il s'agit de la troisième plus grande entreprise de distribution alimentaire de détail aux États-Unis (après Kroger et Safeway). La société était classée 75ème en 2012 dans la liste Fortune 500. Elle réalise un chiffre d'affaires de 36,1 milliards de dollars (2012), et emploie 130 000 personnes. SuperValu exploite 2505 magasins d'alimentation, 878 pharmacies en magasin, 117 centres de distribution de fuel, et approvisionnent 2.200 magasins affiliés supplémentaires.

SuperValu réalise depuis plusieurs années des analyses des « affinités produits », mais comme elle utilise des programmes SQL classiques, elle se contente de travailler une base contenant 13 semaines de données historiques, et met 4 heures à analyser une catégorie de produits. Insatisfaite de cette situation SuperValu a mis en place une plateforme pilote Aster Data, qui contient des informations concernant 225 magasins, 8 ans d'historique de données de transactions, ce qui représente 15 milliards de paniers, ~367 000 codes produits différents de12 catégories de produits (comme par exemple : alcool, céréales, surgelés/glaces, détergent/lessive, fromage, serviettes en papier, pizza, …).

Le premier cas métier traité, a concerné l'analyse des mouvements des prix des articles et leur impact sur la taille des paniers sur une longue durée (6-8ans), la composition des paniers sur une longue durée (6-8ans), les affinités (départements, produits) sur une longue durée (6-8 ans), les meilleures promotions, etc. En s'appuyant sur une solution Aster SQL-MapReduce et en traitant un historique de données de 8 ans, l'analyse des « affinités produits » de toutes les catégories vs tous les autres (11 x 11 catégories), a été réalisée en 48 minutes.

Un autre cas a été l'analyse de la « migration » des consommateurs, l'analyse de la baisse des segments de consommateurs sur une longue période, quels articles ne sont plus achetés, pendant combien de temps les consommateurs sont classés premium avant de se désengager. Un autre cas a été l'analyse de l'incidence de la concurrence : analyse de l'impact des différents concurrents, impact de l'ouverture d'un magasin concurrent sur la taille des paniers, sur la fidélité des consommateurs à savoir si le nombre de paniers par mois se réduit ? Un autre cas a été l'analyse de l'impact des promotions, est-ce que les promotions changent la durée les affinités produits ? Quand il y a une promotion, les clients s'en tiennent-ils à leur marque traditionnelle, ou passent-ils à d'autres marques, ponctuellement ou dans la durée.

Des données des médias sociaux (non structurées / multi-structurées) ont aussi été analysées, pour cela il a fallu intégrer des données de Facebook par exemple avec les données de transaction existantes, et regarder l'impact sur la fidélité des consommateurs, le nombre de fans SuperValu selon les caractéristiques démographiques, voir si les meilleurs consommateurs (Platinum et Gold) sont fans Facebook, etc.

Les résultats ont convaincu SuperValu de l'intérêt de s'équiper pour traiter des Big Data, c'est-à-dire concrètement de pouvoir stocker et analyser toutes sortes de données structurées (comme de longs historiques de transactions), ou comme des données multi-structurées émanant des réseaux sociaux ou de la navigation sur le web. Au final SuperValu a opté

pour un écosystème composite comprenant un entrepôt de données d'entreprise Teradata, un système Hadoop et une plateforme analytique Aster Data, le tout géré par même système d'administration intégré de Teradata.

Aujourd'hui les différents types d'utilisateurs de SuperValu peuvent, suivant leur autorisation, accéder aux différents types de données gérés dans les systèmes et les croiser entre elles librement en utilisant leurs outils habituels préférés. Pour aller plus loin sur ce cas vous pouvez utilement consulter l'interview (7') de Venugopal Adooparambil Senior. IT Manager, EDW & Big Data chez SuperValu : http://www.youtube.com/watch?v=KF91UzmaVHc

27 - Razorfish analyse des big data et crée des expériences clients profitables

Razorfish Inc. est l'une des plus grandes agences de marketing interactif du monde, elle fait partie du groupe Publicis. L'agence fournit différents types de services, du développement web, de la planification et de l'achat de médias, de la publicité utilisant toutes sortes de médias, notamment le mobile, ou les médias émergents, des actions de marketing social, etc. ainsi que des études et des analyses d'impact des programmes de communication.

Razorfish a plus de 2.000 employés dans le monde, avec des bureaux aux États-Unis à New York, Chicago, Seattle, San Francisco, Philadelphie, Portland, Los Angeles, Atlanta, et Austin. En 2005-2007, elle s'est implantée internationalement grâce à des acquisitions à Londres, Paris, Sydney, Hong Kong, Shanghai, Pékin, Berlin, Francfort, Singapour et une joint-venture à Tokyo. Il s'agit donc d'une agence numérique mondiale qui offre un ensemble très complet de services. Selon Forrester c'est le chef de file dans la conception web et le marketing numérique, et une des plus grandes sociétés de marketing et de technologies interactives, capable de créer des expériences client qui permettent de développer les ventes.

Razorfish part du constat que le marketing aujourd'hui doit prendre en compte un contexte exigeant où la concurrence est souvent mondial, où le client omni-canal a des attentes élevées, et sait utiliser à son profit une information abondante sur les offres, en particulier via les sites comparatifs et les appréciations des consommateurs. Une de ses études montre que les entreprises sont moins performantes que l'on croit, moins d'une sur deux sait reconnaître un client numérique qui revient, et donc le traiter différemment d'un prospect. Seulement 12% des entreprises savent reconnaître lors d'une interaction web à quel segment appartient le visiteur.

De plus de trop nombreuses entreprises travaillent en se fondant sur une vision simpliste du parcours du client/prospect avant son achat, or les interactions des personnes avec les entreprises sont nombreuses et très diversifiées : publicité, recherche sur le web, passage magasin, visite du site web, e-mail reçu, sites comparatifs, TV, médias sociaux, centre d'appels, marketing direct, etc. Plus nous savons de chose sur le client/prospect, sur son comportement, sur ses achats passés, plus on peut faire intelligemment des offres, plus on peut pousser des informations pertinentes, plus on peut anticiper ses attentes par comparaison à ses « sosies », qui sont ou qui agissent comme lui.

Pour que les entreprises améliorent leurs résultats, il faut qu'elles arrivent à répondre au mieux aux attentes de chacun de leur client/prospect. Pour cela Razorfish pense qu'elles doivent changer d'approche. Elles doivent investir dans l'étude, la connaissance de leurs clients, en capturant autant que possible toutes les interactions avec eux, en analysant les performances de leurs initiatives de gestion de la relation commerciale en cours, en créant des profils clients tenant compte des achats, des comportements vis-à-vis des canaux, etc. Ceci étant fait, il faut alors définir des stratégies d'expérience client différentes, auxquelles faire correspondre des plans de communication pour délivrer un message au client adapté selon le moment et le canal.

Une telle approche ne peut être mise en place que progressivement. Il faut commencer par la stratégie de marquage et définir un suivi global qui permet d'identifier la clientèle à travers les canaux. Il faut ensuite capturer les données clés pour le marketing & la vente (données en provenance de sites tiers, médias sociaux, données de géolocalisation, serveur de

publicité intégrée à la GRC, données traditionnelles hors web, données de fidélisation, données navigation sur le site, etc.). Il faut alors exploiter ces données (analyse de chemin, analyse d'impact, association, régression, etc.), pour en tirer des informations, une meilleure connaissance des clients et enrichir les capacités de ciblage (sémantique, comportemental, contextuel, démographique, géographique, site, tranche horaire, affinités produits, etc.). Au final il s'agit d'être à même de délivrer le bon message en fonction du moment et de l'endroit : serveur de publicité, ad exchange, site client, email, centre d'appels, médias sociaux, etc.

Pour mettre en œuvre de telles approches, Rasorfish sait traiter les big data nécessaires, notamment toutes celles issues de la traque des clients/prospects sur le web. L'agence s'appuie aussi sur tout un ensemble d'outils marketing classiques, analyse, segmentation, campagne marketing, etc., et demande à ses clients de pouvoir croiser ses informations avec celles d'autres fonctions de l'entreprise (vente, logistique, production, r&d, finance, HR, etc.). Pour analyser ses big data, Razorfish utilise la plateforme de Teradata Aster et ses solutions SQL- MapReduce.

Pour aller plus loin sur ce sujet, vous pouvez utilement écouter l'interview (3') de Matthew Comstock VP Business Intelligence, chez Razorfish : http://www.teradata.com/videos/Razorfish-Finding-New-Insights-with-Data-Driven-Analytics/

28 - De l'expérience big data de Gilt groupe

Gilt Groupe est une entreprise américaine qui collectionne des sites Web proposant à des acheteurs volontaires des ventes flash de marchandises. Elle offre à ses membres des rabais exclusifs sur des vêtements haut de gamme, des chaussures et divers accessoires. Une fois que les membres s'inscrivent, ils reçoivent des alertes par email qui offrent des produits de marque à des prix fortement réduits. Les offres sont majoritairement ciblées en fonction des goûts qui ont été révélés par les achats antérieurs. Les propositions sont valables dans une fenêtre comprise entre 36 et 48 heures. Les offres sont renouvelées tous les jours.

Gilt initialement ciblait les femmes mais a depuis ajouté un site d'habillement pour les hommes, Park & Bond, ainsi que des offres quotidiennes à partir du site Gilt City. Pour attirer les clients, Gilt propose environ 30 différentes offres chaque jour. En suivant constamment la dynamique de ses ventes l'entreprise adapte ses assortiments très rapidement. Ainsi alors que la plupart des grands magasins revoient leur assortiment deux ou trois fois par an, Gilt revoit le sien de huit à 10 fois par an.

Créée il y six ans, Gilt Groupe n'avait au départ aucun système d'information décisionnel, pas d'entrepôt de données, ni de système de reporting, ses données étaient pour la plupart dans des bases opérationnelles fortement dé-normalisées et les données sur le parcours des clients sur les sites étaient dans des fichiers plats. Malgré la conservation de T0 de données (structurées ou non structurées), peu d'analyses étaient réalisées, les requêtes étaient lentes et le montage des opérations commerciales en était affecté. Au fil du temps la situation se dégradait d'autant plus que l'activité se développait et générait 2 millions de lignes supplémentaires chaque jour.

En outre, Gilt Groupe voulait progresser dans la personnalisation de ses offres, améliorer les expériences des clients sur les sites, augmenter les ventes croisées, mieux détecter les fraudes, disposer de statistiques pertinentes sur ses activités, sur ses campagnes marketing. Tout cela supposait de mieux intégrer les données clients, les analyser de façon approfondie de façon à comprendre le comportement des individus, et mettre en lumière des tendances du marché.

Pour être capable de s'adapter à des pointes de charges très importantes, et échapper à d'énormes investissements informatiques, Gilt Groupe s'est tournée vers des solutions Cloud Computing d'Amazon. Ainsi la société dispose d'un système de support client qui personnalise les site ou les offres sur la base de données historiques, et d'un système décisionnel d'une dizaine de nœuds Teradata Aster intégrant toutes ses données auparavant dispersées (achat, navigation, …), compléter de textes en provenance de sources comme twitter, facebook ou les emails entrants, dont les analyses des sentiments

exprimés permettent d'affiner la personnalisation des offres.

Grâce à ses nouveaux moyens Gilt Groupe a gagné un million de nouveaux clients en un an, notamment du fait de la possibilité d'identifier plus facilement des petits segments de personnes qui partagent des intérêts communs pour lesquels des offres spécifiques peuvent être bâties ou mieux profiler les personnes susceptibles d'être fortement intéressées par certaines offres complémentaires. Pour tout cela les informations en provenance de twitter, facebook, etc. furent très intéressantes, en particulier pour toutes les actions de Marketing viral qui sont importante pour l'activité de Gilt Groupe.

En résumé, grâce la plate-forme Teradata Aster, Gilt Groupe relie l'ensemble des données de navigation de ses clients (bannières cliquées, parcours, etc.) avec ses données opérationnelles (vente, achat, etc.), les textes qu'elle peut collecter (web social, email, …), afin d'identifier ce qui se passe chez ses clients, d'optimiser ses assortiments, de définir ses offres promotionnelles et de personnaliser ses propositions. Tout cela permet à Gilt Groupe de suivre le pouls de ses clients et d'être extrêmement réactif.

Pour aller plus loin sur ce cas, vous pouvez utilement regarder l'interview de Geoffrey Guerdat, Director of Data Ingineering de Gilt Group :
http://www.asterdata.com/gilt_groupe_video.php

29 - Les aventures de Wells Fargo dans le big data

Wells Fargo & Company qui a été fondée en 1852, est une des plus anciennes sociétés aux États-Unis. C'est aujourd'hui une société de services financiers diversifiée, qui fournit des services bancaires, des assurances, des prêts hypothécaires, des prêts à la consommation, du financement commercial en Amérique du Nord et à l'étranger. L'entreprise est la quatrième banque aux Etats Unis d'Amérique dans le classement par les actifs (23ème dans le monde), et la deuxième en valeur de marché.

L'objectif de Wells Fargo : « Satisfaire les besoins financiers de ses clients et les aider à réussir financièrement ». En 1995 la société a été la première banque à donner à ses clients l'accès à leurs comptes bancaires via Internet. Aujourd'hui 21 millions de clients de la banque sont actifs en ligne, et génèrent 220 millions de sessions par mois. C'est le canal le plus utilisé par les clients.

Pour supporter ses 84 secteurs d'activité (banque de distribution aux particuliers, services bancaires aux entreprises, prêts, courtage, gestion des risques, traque de la fraude & du blanchiment, …), Wells Fargo a historiquement beaucoup investi dans ses systèmes d'information décisionnels. La société dispose de 4 équipes de support (BI, Data Mining, Entrepôt de données) qui gèrent 6 grands entrepôts d'entreprise (900 T0 au global), beaucoup de petits entrepôts départementaux (de 1 à 10 T0 chacun), de nombreux environnements de fouille de données (300 T0 au global) et d'une solution pour l'analyse des parcours sur le web (12T0).

Dans ce contexte Wells Fargo a essayé de combiner dans des analyses des données structurées et non structurées, et a pu rapidement constater qu'avec ses moyens et méthodes traditionnelles, c'était extrêmement difficile et coûteux. Par exemple le développement d'une application impliquant une demi-douzaine de sources de données structurées et non structurées a été dimensionné à plus de 1700 heures. En effet les données non structurées sont récupérées dans une grande variété de formats, elles doivent donc être d'abord «transformées» pour pouvoir être rapprochées de données structurées, être accessible pour l'analyse, et même dans ce cas SQL peut être extrêmement inefficace en tant que langage d'accès. Wells Fargo en a conclu que sans évolution de la panoplie de moyens décisionnels, ce type d'analyse sera rarement, ou jamais, effectué, l'effort et le coût étant trop élevés.

Pour s'affranchir de cette contrainte, Wells Fargo a décidé de s'équiper d'une plateforme Teradata Aster, pour dans un premier temps prendre en charge les données nécessaires à l'analyse du cheminent des clients sur le web et au reporting correspondant (impact de la recherche payante, des moteurs de recherche, etc…). Cette application a été développée en 370 heures, (320 heures pour développer et coder les structures sous-jacentes et 50 heures pour la

version initiale du rapport d'analyse de cheminement).

Dans un deuxième temps la plateforme a accueilli d'autres données et applications, mais a été réservée à de petits groupes d'utilisateurs expérimentés. La solution Aster avec son cadre MapReduce SQL, permet aux équipes analytiques de mettre en œuvre des fonctionnalités de base de données relationnelle avec de puissantes fonctions d'accès aux données non structurées. Les analystes avancés, les développeurs ETL et DBA sont, pour des opérations basiques, généralement immédiatement productifs et deviennent indépendants après quelques jours de formation à cet environnement.

Depuis ses premières expérimentations, Wells Fargo a beaucoup progressé en matière d'analyse de big data. La plateforme Aster a été augmentée, et a été complétée avec une plateforme Hadoop. De nombreuses applications ont été développées, actuellement les plus gros développements concernent les thèmes de la fraude à la carte bancaire, de l'anticipation de l'attrition et du « marketing attribution ». Pour ces applications de nombreuses données ont été rapprochées (profils des clients, correspondances et courriels, centre d'appels, comptes rendus de visites, activités DAB/GAB, activités web, questionnaires de satisfaction, campagnes de marketing direct ou de publicité), cet ensemble de données offre des possibilités d'analyse qui vont donner du travail aux équipes de support BI pour plusieurs années.

Pour aller plus loin sur ce sujet vous pouvez utilement suivre la vidéo d'une des sessions de la conférence « TiEcon 2013 » durant laquelle des représentants de Wells Fargo, Disney Interactive, Netflix, et Palantir témoignent de leurs expériences en matière de big data.
https://www.youtube.com/watch?v=MA04fYJRQsk

3 - Big data : au-delà des thèmes métiers et des premiers cas d'utilisation

Le traitement des big data n'amène pas de révolution dans le monde du décisionnel, mais élargit le champ de travail des experts de ce domaine, conduit à modifier les infrastructures en place pour répondre à de nouvelles exigences de volume, de variété et de vitesse, à modifier la façon de préparer les données pour réaliser des analyse de pointe, et comme souvent quand le marché aborde un nouveau sujet, des solutions occupent le devant de la scène, même si elles ne sont pas des panacées universelles. C'est le cas d'Hadoop.

Les big data permettent sans aucun doute de mieux connaître les clients, par exemple grâce au traitement automatique de ce qu'ils disent, de mieux les servir via des solutions de commerce électronique et cloud computing, qui sont à la fois de plus en plus sophistiquées et de plus en plus facile à mettre en œuvre.

Mais au final, le big data n'est qu'un prolongement de ce qui se fait depuis des années, et présentent les mêmes avantages, inconvénients ou risques : information, propagande, désinformation et mystification.

31 - Big data : un nouveau champ de travail pour les experts du décisionnel

Depuis quelques temps le mot big data est apparu et est largement utilisé par les journalistes, les analystes, les consultants et certains éditeurs ou constructeurs qui s'intéressent au monde du décisionnel. Mais force est de constater qu'aucune définition ne s'est imposée et que les propos des uns et des autres mélangent allègrement beaucoup de choses : la volumétrie globale des données à traiter, le volume de la donnée de base (web log, texte, photo, vidéo, ...), les types de données (structurées, non structurées, multi-structurées, ...), les ambitions analytiques (aller au-delà de la BI), etc.

Il est difficile de dire si le mot big data est adapté et perdurera, mais il est certain que le domaine du décisionnel connaît un fort développement lié à l'apparition simultanée (quid de la poule et de l'œuf) de « nouvelles » demandes et de « nouvelles » solutions technologiques, qui amènent à traiter de plus en plus de données à la fois en termes de volumes et de variétés. Le volume de données générées quotidiennement dans les systèmes d'information croît de façon exponentielle et donc la volumétrie explose aussi dans le monde du décisionnel. Il y a dix ans Teradata animait un club composés de ses clients qui avaient plus d'un Téra de données dans leur système décisionnel, aujourd'hui nous avons un club pour les entreprises qui ont plus d'un Péta.

Ce qui me paraît plus important dans tout cela, ce n'est pas forcément le volume mais la volonté de développer les approches analytiques avancées en traitant toutes sortes de données brutes, qui nécessitent beaucoup de travail pour en tirer une information métier. Il ne s'agit plus simplement de prendre quelques lignes de factures et de faire quelques opérations basiques pour générer une donnée plus ou moins agrégée qui est du sens. Il s'agit par exemple d'identifier un client/prospect surfant sur le web, de définir s'il a une vision positive ou négative de la marque ou des produits, il peut s'agir aussi de repérer des réseaux d'amis ou de fraudeurs suivant les préoccupations métiers, de mettre en place des moteurs de recommandations qui tiennent compte de votre navigation sur le web et de votre profil, etc.

Pour arriver à disposer des possibilités évoquées ci-dessus, il convient de mettre en place un ensemble de moyens qui permettent d'exploiter des données brutes (par exemple des web logs, des textes, ...), pour en tirer des connaissances (profil, segment, affinité, ...), pour prévoir (attrition, propension, ...) et pour agir (recommandations d'offres, tarification, ...). Concrètement aujourd'hui il faut pouvoir croiser des données classiques d'un système d'information décisionnel, avec des données extraites et structurées via des programmes types MapReduce (Hadoop, Aster Data, ...), et au-delà de l'intégration de ces données, de les travailler avec des outils de data mining pour faire des analyses avancées, mettre au

point des modèles et les utiliser pour enrichir des processus opérationnels, par exemple au niveau des sites web, des centres d'appels, des divers canaux d'interaction avec le client.

Teradata qui œuvre depuis longtemps dans l'analyse de données ne pouvait pas passer à côté de ce nouveau domaine prometteur. Comme les solutions classiques sont mal adaptées pour faire certains traitements nécessaires pour ces données « nouvelles » pour le monde du décisionnel, Teradata a donc racheté Aster Data au début de l'année 2011, une société spécialisée offrant une solution brevetée SQL-MapReduce™. Avec ces moyens complémentaires permettant de mieux exploiter des volumes importants de données non relationnelles, Teradata va pouvoir proposer plus de solutions analytiques innovantes à ses clients cherchant à utiliser leur système d'information pour différencier leur positionnement sur le marché (analyse des relations clients et des réseaux, optimisation du marketing, détection et prévention des fraudes, etc.).

Pour aller plus loin sur la vision Teradata du Big Data et les solutions Aster Data vous pouvez utilement consulter les sites suivants : www.asterdata.com & www.teradata.com

32 - Infrastructure big data : répondre à des exigences de volume, de variété et de vitesse

Un système décisionnel big data est reconnu comme tel, s'il présente des capacités particulières en matière de volume, de variété de données et de vitesse de traitement.

Aujourd'hui les entreprises pour améliorer leurs modèles de connaissances et de prévisions, n'hésitent pas prendre en compte plusieurs centaines de facteurs, et pour cela mettent en place de nouveaux moyens d'analyse qui permettent de traiter de grands volumes de données. Or le traitement de grands volumes de données est un défi pour les infrastructures décisionnelles habituelles. Stocker de grands volumes n'est pas un problème, mais les exploiter nécessite des architectures massivement parallèles, des entrepôts de données tels que ceux proposés par Teradata par exemple, ou des « solutions MapReduce » telles que celles d'Hadoop ou d'Aster Data. Ici le choix de la solution dépend de la variété des types de données à traiter et de la vélocité attendue. En effet MapReduce est meilleur qu'une base de données relationnelle pour traiter des données non structurées, et Hadoop est batch alors qu'Aster Data est temps réel. Comme il n'y a pas de solution miracle, les grandes entreprises se dotent d'un mix de moyens leur permettant de bénéficier des avantages des différents types de solutions.

A partir du moment où l'on veut prendre en compte toutes sortes de données, des textes, des données issues de capteurs divers, des données de géolocalisation, des données de réseaux sociaux, des images, etc..., ces données ne se présentent pas sous une forme parfaitement ordonnée et ne sont pas d'emblée prêtes pour une exploitation analytique. Même les données issues du web ne se sont pas dès le départ parfaites. Une tâche courante des systèmes Big Data est de prendre en charge des données non ou multi-structurées et de les traiter pour les rendre consommables par des humains ou des applications analytiques. Un exemple classique en matière de traitement de textes est de déterminer à quoi réfère un mot : Paris est-ce la capitale de la France ? La ville Paris dans l'Illinois ? Paris la célèbre people ? Etc. Il s'agit aussi de stocker de la façon la plus performante possible des données, et les bases de données relationnelles ne sont pas toujours la meilleure solution, par exemple pour les données XML ou pour les réseaux de relations qui sont des graphiques. Même là où il n'y a pas une incompatibilité de type de données, un inconvénient de la base de données relationnelle est le caractère statique de ses schémas. Les bases de données semi-structurées NoSQL fournissent assez de structure pour organiser les données, mais ne nécessitent pas un schéma exact des données avant de les ranger.

Les exigences de vitesse de traitement des données ont ces dernières années augmentées de façon analogue à celles des volumes. Cela ne concerne plus seulement quelques sociétés spécialisées comme les opérateurs financiers (les traders), mais touche la plupart des secteurs économiques. A l'ère d'internet et des mobiles le rythme des affaires s'est accéléré, nous ne consommons plus de la même façon, les

formes de concurrence ont évoluées et les flux d'information aussi. Par exemple les détaillants en ligne sont en mesure de suivre les clics de chaque client, de leur première interaction à la vente finale. Ceux qui sont capables de rapidement utiliser cette information, en recommandant des achats supplémentaires par exemple, acquiert un avantage concurrentiel notable.

Le défi ne tient pas uniquement dans le fait de devoir assumer le volume ou la vitesse des données entrantes, mais surtout dans la vitesse des analyses et du déclenchement des actions pertinentes. La fraîcheur des informations délivrées est primordiale. Par exemple : Traverseriez-vous une rue sans regarder, en vous fiant à une vue de la circulation prise cinq minutes avant ? La vitesse de rétroaction est une source d'avantages concurrentiels, en particulier pour toutes les activités web. Face à de tels besoins les technologies habituelles du monde du décisionnel sont dépassées par le rythme, et seul un mix de solutions permet de répondre aux attentes métiers. C'est ainsi que des clients Teradata comme eBay ou LinkedIn par exemple, utilisent à la fois des systèmes Teradata (Active Entreprise Data Warehouse, Extreme Data appliance, Extreme Performance Appliance) et des systèmes Hadoop ou Teradata Aster.

Pour aller plus loin sur la vision de Teradata des infrastructures big data, vous pouvez venir nous écouter lors du Congrès big data Paris 20 & 21 Mars 2012 : http://www.bigdataparis.com/fr-index.php

33 - De la préparation des big data pour les analyses avancées

Les professionnels de la BI et de l'entrepôt de données sont convaincus que les données qui alimentent les rapports doivent être de qualité, intégrées et documentées. Pour répondre à ces exigences, les équipes travaillent dur pour définir un modèle, extraire, transformer et charger, des données de qualité, gérer les données de référence et les métadonnées. Au-delà du souhait des décideurs de disposer des meilleures données possibles pour fonder leurs décisions, les entreprises doivent aussi tenir compte du fait que certaines données sont publiques, et que les erreurs qui pourraient les entacher pourraient être désastreuses pour l'entreprise.

La préparation des big data pour les analyses avancées se fait dans un contexte très différent. Ces données ne sont jamais rendues publiques et ces analyses sont souvent ponctuelles ou très rarement réitérées de la même façon. Par conséquent la modélisation et la gestion des chargements, de la qualité etc. ne se fait pas avec les mêmes exigences. En fait, si vous appliquez l'arsenal complet des pratiques de préparation des données sur des données analytiques, vous courez le risque d'en réduire leur valeur analytique.

Comment un processus de préparation sensé donner de la valeur aux données, peut-il nuire ? Pour répondre à cette question, voyons d'abord ce qu'on appelle «analyse avancée ». Ces techniques analytiques seraient mieux appelées «analyses exploratoires», car c'est ce que les utilisateurs font avec elles. Des professionnels de l'analyse ou des analystes métier utilisent ces techniques impliquant la mise en œuvre de programmes SQL complexes ou de MapReduce, pour explorer des données et découvrir des faits que l'on ne connaissait pas auparavant. Par exemple découvrir un ensemble de transactions qui indiquent un nouveau type de fraude, ou un nouveau groupe de clients ayant un comportement homogène, ou un groupe de caractéristiques possédées par les personnes qui passent à la concurrence.

Généralement, vous ne pouvez pas faire ce type de découverte à partir des données modélisées, agrégées et déjà excessivement étudiées de votre entrepôt d'entreprise. Pour cela vous avez besoin de big data, beaucoup plus détaillées telles qu'elles sont dans leur système source, certaines formes d'analyse s'accordant bien à des données brutes, apparemment incomplètes. Par exemple, l'efficacité d'applications analytiques pour la détection de fraudes peut dépendre de valeurs aberrantes, de données non-standard ou de données manquantes, pour indiquer la possibilité d'une fraude.

Les possibilités de découverte se concentrent souvent sur un tout petit nombre de clients, de transactions, sur une période de temps très courte, etc. Ces

tranches fines peuvent facilement disparaître dans une passe d'agrégation. Ainsi, si vous appliquez les processus habituels d'extraction, de transformation et de chargement de données ou ceux liés à vos exigences de qualité, comme cela se fait aujourd'hui pour un entrepôt de données classique, vous courez le risque d'éliminer les pépites qui font des big data un trésor pour la découverte de nouveaux aspects de vos affaires. C'est pourquoi la préparation des big data semble minime (même bâclée) - souvent juste des extraits et des jointures de tables - par rapport à la gamme complète des préparations appliquées aux données d'un entrepôt d'entreprise.

Est-ce à dire que nous pouvons jeter les meilleures pratiques en matière d'ETL, de Qualité, de Métadonnées, de MDM et de Modélisation des données ? Non, bien sûr que non. Après que les experts techniques et métiers aient fait les premières analyses sur leur big data, ils ont généralement besoin pour exploiter complétement ce qu'ils ont découverts, de rapprocher leurs résultats avec des données de l'entrepôt d'entreprise pour enrichir les référentiels et les analyses métiers (BI ou data mining). Par exemple, lorsque l'analyse de big data révèle de nouveaux éléments métiers clés - comme de nouvelles formes de désabonnement, des segments de clientèle, des coûts induits, ... – ces connaissances doivent être intégrées dans l'entrepôt et dans les rapports, afin que les décideurs puissent en tirer profit

Pour aller plus loin sur ce sujet vous pouvez participer au CONGRÈS BIG DATA PARIS : ENTREZ DANS L'ÈRE DU DÉLUGE DE DONNÉES que Teradata Aster sponsorise : www.bigdataparis.com

34 - Hadoop n'est pas la panacée universelle

Les études des experts du marché de l'IT, indiquent que les analyses avancées de données variées et la gestion de volumes croissants de données, sont parmi les cinq priorités des directions informatiques des grandes entreprises et des sociétés en pointe dans le monde de l'internet. Ainsi le big data prend de plus en plus d'importance, et un nombre croissant d'entreprises complètent leur infrastructure décisionnelle, avec de nouvelles plates-formes analytiques pour améliorer leur efficacité et leur rentabilité.

L'observation des premières expériences big data montre que de nombreuses technologies différentes sont utilisées, même si une technologie émergente, basée sur le projet open-source Apache Hadoop, est très souvent présente dans les infrastructures des pionniers du big data. La popularité d'Hadoop semble résider dans sa capacité à traiter de grands volumes de données, avec une infrastructure faite de grappes de serveurs standards low cost. Mais attention tous les experts indiquent qu'il n'y a pas de solutions universelles en matière de big data, que les utilisateurs doivent déterminer en fonction de leurs besoins le mixte de technologie qu'il convient qu'ils mettent en place, et définir précisément où chacune (dont Hadoop) peut ajouter de la valeur dans leur architecture décisionnelle.

D'abord, il faut considérer les données qui sont à traiter. Il y a celles dont le modèle de données est établi et stable dans le temps. Ici nous trouverons tout ce qui concerne la BI classique, le reporting, les analyses financières, les décisions automatisées liées à l'opérationnel, et l'analyse des données spatiales. Il y a celles que l'on peut stocker de façon plus ou moins brute, et qui vont être modélisées de différentes façons suivant les besoins des analyses itératives. Cela peut concerner par exemple les analyses des clics des utilisateurs dans leur navigation sur le web, les données des capteurs, les CDR dans les télécommunications. Il y a celles qui sont simplement définie par un format. Il s'agit par exemple des images, des vidéos, et des enregistrements audio.

Il faut aussi considérer ce que l'on veut faire des données. Si l'idée est d'exploiter classiquement (reporting, BI, data mining) des données structurées, un appliance classique convient parfaitement, on peut éventuellement le compléter d'un stockage bon marché avec une solution Hadoop ou un appliance spécifique comme « l'Extreme Data Appliance » de Teradata pour certaines données qui n'ont pas grand intérêt à être intégrées dans le modèle de l'entrepôt d'entreprise. Pour les autres données (web log, capteurs, CDR, images, vidéos, ...) il faut avoir

recours suivant les traitements prévus à des solutions du type Hadoop et/ou Teradata Aster, qui permettent de mettre en œuvre à moindre coût (stockage, développement d'application, exploitation, intégration avec l'entrepôt de données structurées) des programmes MapReduce.

A noter que la solution Teradata Aster MapReduce utilise une technologie brevetée SQL-MapReduce qui permet de mettre en œuvre des programmes MapReduce sans avoir à apprendre un nouveau langage de programmation. Cette solution offre aussi des performances, une évolutivité pour prendre en charge de gros volumes de données, et traiter des données relationnelles avec des données relevant de divers formats. Par rapport à Hadoop cette solution offre des avantages conséquents en matière de charges de développement d'applications et de temps de réponse des requêtes par exemple.

Qu'elle soit mesurée par une augmentation des revenus, des gains de parts de marché ou la réduction des coûts, l'analyse des données a toujours joué un rôle clé dans la réussite des entreprises. Aujourd'hui le développement d'internet et de processus d'affaires automatisés, rend crucial l'exploitation des Big Data, et amènent les dirigeants d'entreprises à dépendre de plus en plus de leurs moyens d'analyse de données. Dans ce contexte, les équipes informatiques sont alors amenées à compléter leurs infrastructures décisionnelles existantes, avec de nouvelles solutions qui permettent de mettre en œuvre des algorithmes complexes. Les pionniers qui ont déjà exploités des big data avec succès disent tous qu'il n'existe pas de solution miracle, pas plus Hadoop qu'une autre, c'est pourquoi Teradata propose différentes plateformes mettant en œuvre la base de données Teradata, la solution Aster MapReduce ansi qu'Hadoop.

Pour aller plus loin sur le sujet vous pouvez utilement découvrir le dernier Teradata Aster Big Analytics Appliance qui intègre dans une même plateforme les solutions Aster et Hadoop (distribution Hortonworks) :
http://www.asterdata.com/resources/downloads/datasheets/Teradata-Aster-Big-Analytics-Appliance-Datasheet.pdf

35 - Des big data pour mieux servir les clients

Dans le contexte actuel de concurrence exacerbée, les grandes entreprises doivent, pour se maintenir au plus haut niveau, jouer de toute une palette de moyens : engagements de service, approche multicanal, contrôle poussé de la chaine logistique jusqu'au dernier kilomètre, communication digitale, magasin connecté, site internet, centre d'appels, programme de fidélisation, animation de communauté, personnalisation des produits et des prix etc. Mais par-dessus tout elles doivent constamment placer le client au cœur de leur stratégie, et faire qu'il soit au premier rang des priorités de tous leurs collaborateurs, qui doivent faire du sourire une composante clé de leur métier.

Force est de constater que beaucoup d'entreprise ne font pas vivre à leur client que de bonnes expériences : files d'attente longues (quelles soient virtuelles au téléphone ou concrètes en magasin), vendeurs plus ou moins compétents et parfois sans scrupules, processus de prise de commande par téléphone ou internet compliqués et longs, clauses contractuelles opaques, engagements non tenus, services client injoignables, réclamations non prises en compte, au moindre problème spirale infernale d'appels inutiles, absence d'historique des échanges, etc.

Au-delà de la définition d'une stratégie focalisée sur le client, de l'adaptation de l'organisation, de la formation des collaborateurs et de la mise en place de systèmes d'information performants pour gérer les opérations, les grandes entreprises qui veulent faire la course en tête doivent beaucoup investir dans leurs systèmes décisionnels, et en particulier utiliser les Big Data pour mieux comprendre les marchés, maîtriser leurs actions et faire vivre à leurs clients les expériences qu'ils attendent. Mais attention il n'existe pas de solutions toutes faites, prêtes à l'emploi, qui permettent de résoudre tous les problèmes préalablement cités.

Même si les technologies des systèmes décisionnels traditionnels continuent d'évoluer, elles sont dans l'ensemble très matures et les professionnels savent les mettre en œuvre efficacement. Il en va tout

autrement pour les technologies liées aux big data qui sont pour la plupart d'une diffusion très récente, en devenir, mal connues et mal maîtrisées par le petit nombre de professionnels qui ont un peu d'expérience dans le big data. Malgré ce contexte peu favorable, il est nécessaire pour les grandes entreprises de lancer des POC pour commencer à s'initier aux big data dont le traitement demain sera incontournable.

Si pratiquement toutes les fonctions de l'entreprise peuvent être concernées par les big data, actuellement c'est surtout la fonction Marketing / Vente qui s'en préoccupe. Pour elle il s'agit principalement de trouver de nouvelles opportunités, à travers la traque des usages et des expériences des clients, de mettre en lumière des attentes, des besoins, auxquels on peut répondre avec des offres innovantes, d'exploiter au mieux les relations, de savoir positionner, pousser l'offre auprès des personnes à potentiel en exploitant au mieux les moments de contact, d'optimiser leurs investissements marketing en fonction des résultats qu'ils ont pu observer de leurs programmes antérieurs.

Mais attention si la fonction Marketing / Vente offre beaucoup de possibilités, il n'est pas toujours simple, pour une entreprise déjà bien équipée en systèmes d'information décisionnels, de trouver par où commencer avec les big data. En effet dans certain cas d'usage, les big data ne font qu'apporter un peu plus d'information pour étayer une analyse, pour améliorer les résultats d'un modèle existant, et donc souvent la nouvelle application big data , bien qu'elle offre un apport significatif, n'a pas un ROI très intéressant. Il est donc préférable de chercher à investir dans des champs nouveaux d'analyse qui sont peu ou pas couverts, et ne revenir à l'amélioration des pratiques existantes que dans un deuxième temps.

Pour trouver les bons champs à investir, on peut par une bonne approche d'intelligence économique s'inspirer des pionniers, des très grandes entreprises internationales qui font la course en tête. Teradata par exemple, a créé un club des entreprises qui ont plus de 1 P0 dans leur système décisionnel et utilisent de nouvelles technologies big data. Ce club compte actuellement près d'une quarantaine d'entreprises, et des représentants dans presque tous les principaux secteurs d'industrie (eBay, Apple, Boeing, Wal Mart, Barclays, AT&T, Pfizer, Well Point Health, Intel, etc.). On peut aussi s'inspirer de plus petites entreprises qui n'ont pas 1 P0, mais pour lesquelles les big data sont vitales comme : LinkedIn, Gilt Groupe, Chegg.com, Zazzle, Eightfoldlogic, Razorfish, Insight Express, Machinima, Mzinga, Intuit, Full Tilt Poker, etc.

Pour aller plus loin sur ce sujet vous pouvez utilement consulter des cas clients Teradata accessibles via les liens suivants : http://www.asterdata.com/customers/index.php et http://www.teradata.com/case-studies/

36 - Big data et traitement automatique du langage naturel

Le Traitement Automatique du Langage Naturel (TALN ou NLP en Anglais pour Natural Langage Processing), est une branche de l'informatique, centrée sur le développement de systèmes qui permettent aux ordinateurs de communiquer avec les humains, en utilisant le langage courant. Le TALN est considéré comme un sous-domaine de l'intelligence artificielle, et a un chevauchement important avec le domaine de la linguistique informatique ou computationnelle.

Concrètement il s'agit : de systèmes de compréhension du langage naturel qui convertissent le langage humain en représentations qui sont plus faciles à manipuler pour les programmes informatiques, ou de systèmes de génération de langage naturel qui convertissent les informations de bases de données informatiques en langage lisible par l'homme. Le TALN concerne à la fois le texte et la parole, mais le travail sur le traitement de la parole a évolué dans un champ distinct.

Pourquoi le TALN, à quoi cela sert-il ? Les applications qui ont à traiter de grandes quantités de textes nécessitent une expertise en TALN. C'est expressément le cas lorsque l'on veut :

- classer des textes en catégories, indexer et mener des recherches dans de grands ensembles de textes (classer les documents par thèmes, langue,

auteur, filtrer les spam, rechercher des informations pertinentes, déterminer les sentiments (positif, négatif),

- extraire des données de textes en convertissant des données non structurés en données structurées,
  extraire des informations, comme par exemple de lister les noms des personnes et des événements auxquels ils participent, à partir d'un document.
- automatiser la production de résumés (condenser 1 livre en 1 page, ...),
- trouver des réponses à des questions en langage naturel dans une collection de texte ou base de données,
  corriger l'orthographe, la grammaire,
- détecter des plagiats,
- traduire automatiquement,
- etc.

Pour les systèmes informatiques la tâche est rude. Quand les humains de 2013 voient un texte, ils le lisent et le comprennent (sous réserve de connaître le langage utilisé), quand les ordinateurs 'voient' un texte, ils ne perçoivent que des chaînes de caractères (ou des balises HTML). Le TALN est difficile, car la langue est souple, il y a constamment de nouveaux mots, de nouvelles significations, des significations différentes dans des contextes différents, la langue est subtile, la langue est complexe, il y a de nombreuses variables cachées (connaissances sur le monde, connaissances sur le contexte, connaissance des techniques de la communication humaine, problème d'échelle, …).

Dans ce domaine Teradata propose des solutions analytiques associant Aster et Attensity, elles permettent de traiter facilement de gros volumes de données textuelles, de les analyser et de leur donner du sens. Concrètement il s'agit de faciliter l'application des principes linguistiques pour extraire du contexte, des entités et des relations, de façon similaires à ce qu'un humain ferait ; faciliter la détection automatique et l'extraction d'entités telles que nom, lieu, … ; faciliter l'utilisation de règles de classification personnalisés pour classer les textes par leur contenu, trier par pertinence, et découvrir des informations. Il s'agit aussi de rapprocher ces données des historiques des transactions ou des contacts, et de comprendre en fonction de ce que les clients ont exprimés sur le web, ce qui ne va pas ou par quoi ils sont intéressés, de définir des communications, des offres appropriées, ou d'identifier des clients, des cibles à fort potentiel.

Pour aller plus loin sur ce sujet vous pouvez utilement consulter le lien ci-dessous :
http://www.teradata.com/partners/Attensity-Group/

37 - Big data, commerce électronique et cloud computing

Le commerce électronique est dépendant d'un ensemble de technologies, d'applications et de processus d'affaires qui lient les entreprises et les consommateurs, pour l'intégration et l'optimisation des opérations au sein et entre les entités participantes, pour l'achat, la vente et la fourniture de produits ou services. Attention cependant à ne pas surestimer l'importance de la technologie, et à ne pas se fier à la trompeuse réputation du commerce électronique, d'exiger un faible coût d'entrée, de ne nécessiter que peu de ressources humaines et d'avantager les premiers installés.
On peut distinguer quatre grands types de modèles d'affaires de commerce électronique. Le distributeur virtuel comme Amazon qui n'utilise que le canal du web et génère la quasi-totalité de son chiffre d'affaires de la vente en ligne. Les « clicks & mortar » comme la FNAC, qui dispose d'un réseau de magasins physiques en tant que canal de distribution principal, mais aussi des offres en ligne. Les vépécistes qui historiquement vendent via un catalogue papier, mais aussi depuis quelques années via un site web. Enfin les fabricants qui vendent directement aux consommateurs, soit parce que c'est leur seul canal de vente (Cf. DELL), soit qu'ils travaillent en parallèle via de multiples canaux.

Le commerce électronique tire ses revenus des marges qu'il réalise de ses opérations commerciales (Amazon, Fnac, Dell), de la publicité (seule source de revenu pour Yahoo par exemple), des frais de transaction facturés (eBay), ou de frais de souscription (Journaux en ligne par exemple). Pour réussir il convient de disposer d'une bonne offre, d'une communication efficace (site web, trafic,

expérience client, …), et pour les biens matériels d'une distribution physique adaptée.

Ce qu'il y a de bien avec le commerce électronique, c'est qu'Internet est un canal très mesurable : combien de visiteurs, combien d'acheteurs, combien de temps ont-ils passés sur le site, quelles pages ont été vues, etc. Cependant identifier les visiteurs n'est pas toujours simple, sauf à les faire s'enregistrer sur le site. En effet une bonne gestion des cookies et des adresses IP n'offre pas une garantie absolue pour les analyses et les mesures, du fait des suppressions régulières des cookies, des changements fréquents d'adresse IP, et du fait que dans tous les cas, on identifie une machine (ou un navigateur), qui peut représenter plusieurs personnes ou une partie seulement de l'activité internet d'une personne.

Pour gérer et mesurer son activité, l'e-commerçant doit relever un certain nombre de défis. Il doit être capable par exemple d'assumer la disponibilité de son site (24/7, 99,99%) et son évolution, de répondre aux pics d'activité dans la journée et dans l'année, d'assurer le stockage d'un grand volume de données brutes, de mettre à disposition des analystes les données d'activité, et de supporter la charge que représente la fouille approfondie des big data conservées. C'est pourquoi beaucoup d'e-commerçant se tournent vers des solutions de cloud privées ou publiques, gérées par des entreprises spécialisées (exemple Amazon EC2), qui louent leurs services (infrastructure, plateforme de développement, applications opérationnelles et décisionnelles).

Des solutions d'analyse avancée, de data mining de big data, comme celles de Teradata Aster, SAS ou Hadoop (Hortonworks, Cloudera) par exemple, sont disponibles pour les e-commerçants, ils peuvent les installer sur leur système ou en disposer en mode cloud,. Comme pour la gestion des sites et des transactions, le cloud offre des avantages en matière de systèmes d'information décisionnels, et de nombreuses sociétés y ont recours, chez les clients d'Aster c'est le cas d'Insight Express, de Mzinga, de Gilt, …. Pour plus d'information sur la solution cloud de Teradata Aster vous pouvez consulter le lien suivant : http://www.asterdata.com/product/deployment/cloud.php

38 - Big data : information, propagande, désinformation & mystification

La vague du big data nous fait franchir une nouvelle étape de l'ère de l'information dans laquelle nous vivons. Les données, les informations ont pris une place centrale dans toutes nos activités, même la guerre se porte sur ce terrain (guerre de l'information, guerre électronique par exemple). A l'ère de l'information, la puissance des idées joue un rôle de plus en plus éminent, mais en faisant du scepticisme un principe, notre époque devient aussi l'ère du soupçon et de la défiance généralisée. Au départ il y a toujours des données qui sont transformées en information, ce qui nécessite en particulier un savoir-faire, des connaissances, une éthique, et nous expose à de nombreuses torsions : secret (information fermée), rumeur (degré zéro de l'information), propagande, désinformation, voire mystification.

La publicité est une forme de propagande qui est acceptée dans notre société, elle est employée pour développer une « plaidoirie » pour un produit ou une marque par exemple. Il s'agit de promouvoir un point de vue, de convaincre, mais aussi de décérébrer, par exemple dans cette optique le rôle de la publicité suggestive est de prendre toute les précautions pour éviter que le cerveau travaille de façon analytique et critique. La publicité, comme toutes les actions de propagande a des objectifs de conditionnement individuel ou des foules. La propagande cherche à séduire, à rallier, à convertir les indécis, à renforcer le camp des convaincus et autres sympathisants, à ébranler les sceptiques, mais aussi dans certains cas (hors publicité) à démoraliser, à terrifier les opposants. Il convient de noter ici que le terrorisme est une action de propagande qui fonde sa puissance sur la violence du message.

La désinformation quant à elle, consiste à volontairement fausser la compréhension d'une situation. Désinformer, c'est fournir de fausses indications afin d'orienter dans une mauvaise direction. Il s'agit le plus souvent de présenter de façon tendancieuse un fait avéré, de fournir une

information biaisée, de brouiller les pistes, de noyer l'information vraie dans une avalanche de considérations et de détails. Les champions toutes catégories de la désinformation sont les politiciens. Les expressions politiquement correctes sont une forme de désinformation, la langue de bois une autre forme. La grande question ici, est de savoir s'il peut exister une désinformation moralement honnête, dès lors qu'elle serait au service de fins respectables ?

En matière de manipulation de l'information, la médaille d'or revient aux actions de mystification. Il existe différentes formes de mystification, des douces (canulars, poissons d'avril, hoax) et des durs (arnaques commerciales par exemple). Les militaires sont les experts de ce type de manipulation, déjà hautement recommandée cinq siècles avant Jésus Christ par Sun Tzu qui consacre un chapitre de son célèbre livre à la tromperie. Les références militaires pour ce type d'action sont très nombreuses : Cheval de Troie, Horace/Curiace, Dépêche d'Ems, et durant la deuxième guerre mondiale, Pearl Harbour, Barbarosa, Mincemeat et Overlord. Recourant systématiquement au stratagème, la mystification suppose un montage complet comportant une combinaison de faux indices, un enchaînement de leurres, l'utilisation de transfuges, d'espions et de traitres afin de conduire l'ennemi qui se méfie à la mauvaise décision et au comportement fatal.

Le monde du big data nous est présenté aujourd'hui sous son angle sympathique d'amélioration de la compréhension des situations, de la connaissance des clients, de la traque des fraudes, …, mais nul ne doute qu'il sera une bonne source pour alimenter de nombreuses actions de propagande, de désinformation et de mystification.

Pour aller plus loin sur le thème de l'utilisation des big data par les entreprises, vous pouvez suivre cette courte vidéo (4') de Marc Parrish, Vice President of Customer Loyalty and Retention at Barnes & Noble, présentant comment ce grand distributeur américain exploite ses big data pour améliorer ses relations clients : http://www.asterdata.com/barnes-and-noble-video.php

Annexes

Pour aller plus loin, vous pouvez consulter mes autres publications sur le sujet.

En Anglais

Des présentations consacrées aux big data :

1. Text Mining - http://www.decideo.fr/bruley/docs/1___text_mining_v0a.ppt
2. Sentiment Analysis - http://www.decideo.fr/bruley/docs/2___sentiment_a_v0.ppt
3. Social Network Analysis – http://www.decideo.fr/bruley/docs/3%20-%20SNA%20V0mb.ppt
4. Web Log & Clickstream - http://www.decideo.fr/bruley/docs/4___web_log_1
5. MapReduce - http://www.decideo.fr/bruley/docs/5___mapreduce_v0.ppt
6. Marketing Attribution - http://www.decideo.fr/bruley/docs/6%20-%20Mkg%20Attribution%20V0.ppt
7. Social CRM - http://www.decideo.fr/bruley/docs/7___scrm_v00.ppt
8. Churn - http://www.decideo.fr/bruley/docs/8___churn_com___v0.ppt
9. Machine Learning - http://www.decideo.fr/bruley/docs/9___machine_learning_v0.pptx
10. Product Affinity - http://www.decideo.fr/bruley/docs/10___product_affinity___v0.ppt
11. Next Best Offer - http://www.decideo.fr/bruley/docs/11___next_best_offer___v0.ppt
12. Visualization - http://www.decideo.fr/bruley/docs/14___visualization___v0.ppt
13. Geomarketing – http://www.decideo.fr/bruley/docs/16%20-%20GeoMKG%20-%20V0.ppt
14. Natural Language Processing - http://www.decideo.fr/bruley/docs/17___nlp___v0.ppt
15. Pricing - http://www.decideo.fr/bruley/docs/18%20-%20Pricing%20V0.ppt

En Français

Vous pouvez aussi consulter mon eBook : ***Propos sur les SI Décisionnels*** qui décrit à quoi ils servent, les bonnes façons de les organiser, leur utilité pour la fonction marketing et les autres fonctions de l'entreprise, la façon de les gérer ainsi que des cas remarquables. Cliquez ici. http://www.decideo.fr/eBook-Propos-sur-les-Systemes-d-Information-Decisionnels_a4507.html

2014