TECHNOLOGIE

François Marcotorchino
(Directeur Scientifique
IBM France)

# Le « Business de l'Information » et ses composantes : *Data Mining, Data Warehouse, Text Mining*

## ► 1 – Introduction

Depuis ces cinq dernières années, des changements profonds s'opèrent dans les relations qu'entretient l'entreprise (qu'elle soit publique ou privée) avec les acteurs du marché : clients, médias, fournisseurs, concurrents, etc. Cet article vise à donner un nouvel éclairage aux raisons à la fois structurelles et méthodologiques relatives à ces changements ainsi que les impacts qui en découleront pour les entreprises.

Des changements profonds s'opèrent dans les relations qu'entretiennent les entreprises avec les acteurs du marché.

Ces changements sont essentiellement liés à la nécessité pour l'entreprise de prendre en compte et d'analyser des volumes de plus en plus importants de données informationnelles et hétérogènes arrivant par des canaux multiples (données historiées de type marketing ou financières, données résultant d'échanges internes / intranet, données externes / internet, centres d'appels, médias, etc.). Pour être exploitées à des fins décisionnelles, ces données doivent être stockées, structurées, filtrées, résumées, diffusées et « réaccédées » de façon performante et optimale.

Ce besoin nouveau d'exploitation des données qui s'est d'abord exprimé aux États-Unis sous le nom de *Business Intelligence,* a profondément modifié la conception et le rôle des services informatiques dans l'entreprise. Loin de signifier la fin de l'informatique décisionnelle traditionnelle en tant que telle, ce nouvel engouement pour une informatique à haute valeur ajoutée va complètement relancer les métiers qui s'y rattachent (informaticien, statisticien, décideur, etc.) en leur donnant un positionnement plus stratégique au sein de l'entreprise d'aujourd'hui. Du fait de l'hétérogénéité des sources et de la nature des données à traiter, ces métiers devront acquérir des compétences pluridisciplinaires tant dans les domaines scientifiques que dans les domaines métiers associés. Ils acquerront de ce fait une plus grande visibilité, ils participeront davantage aux orientations stratégiques de l'entreprise et verront croître leurs responsabilités.

Ces changements résultent de la nécessité pour l'entreprise de prendre en compte et d'analyser des volumes importants de données hétérogènes.

Les nouveaux métiers, dont il est ici question, doivent s'appuyer sur les connaissances théoriques et pratiques enseignées dans des cursus à la fois universitaires et professionnels, tels que l'on peut en rencontrer (malheureusement en trop petit nombre) aujourd'hui en France (ENSAE, ISUP, ENSAI, HEC, DEA des universités de Paris VI, Rennes, Montpellier, Marseille, Marne la Vallée, Orsay, etc.). A titre d'exemple, les enseignements en statistique et mathématiques de la décision seront complétés par un enseignement sur les structures de recueil de l'information que sont les *Data Warehouse* (que nous définirons ultérieurement), ainsi que l'extension souhaitée à des rudiments de linguistique ou d'analyse d'images. Car analyser l'information, pré-

sente dans l'entreprise aujourd'hui, ne se limite plus au seul traitement des données codées ou structurées, apanage de la statistique traditionnelle, mais également à toute information qui lui parvient soit directement (Internet, Intranet) soit indirectement par les supports presse ou média qui en général véhiculent de l'information non structurée et a fortiori non codée (messages du WEB, communiqués de presse, *news* d'agences type AFP ou Reuters, lettres des clients, etc.).

S'appuyant sur l'expérience acquise, tant en milieu industriel qu'en milieu universitaire, il nous est apparu que les entreprises de grande taille et de taille moyenne qui veulent se doter d'un système central efficace d'analyse décisionnelle ont, outre un besoin vital de supports informatiques et logiciels associés à ce domaine, un besoin également crucial de compétences de haut niveau dans ces approches que seuls des profils équivalents à ceux décrits précédemment sont capables de satisfaire.

## ▶ 2 – La génèse du changement

Trois facteurs influencent la mise en place de systèmes d'analyse à but décisionnel : le marché, la technologie et le volume de données et d'informations.

La révolution liée au traitement de l'information qui se produit dans le milieu industriel aujourd'hui est due à un certain nombre de facteurs qui, tous, concourent au développement favorable de la mise en place de systèmes d'analyse de l'information à but décisionnel. Ces facteurs sont au nombre de trois et portent sur des aspects de marché, de technologie et de volume de données et d'informations à traiter.

**Le marché** influe sur l'entreprise à différents niveaux :

- une pression accrue de la concurrence,
- une obligation de se différencier sur les produits et services proposés,
- un marketing plus ciblé sur le client, obligeant de facto à une amélioration drastique de sa connaissance individuelle et de ses aspirations et désirs,
- une réactivité plus importante en termes de choix et de décisions à prendre,
- une compréhension plus nette par les entreprises de ce qu'elles peuvent gagner par une étude et exploitation approfondies des données et informations dont elles disposent.

**La technologie** impacte l'entreprise selon les axes suivants :

- une plus grande puissance des ordinateurs et des systèmes générant des coûts informatiques allant en diminuant à périmètre constant,

- des logiciels adaptés aux traitements de grands ensembles de données et plus simples d'utilisation,
- l'avènement de nouvelles méthodologies statistiques et mathématiques,
- une présentation des résultats d'analyse sous une forme graphique plus conviviale et plus compréhensible de la part des décideurs.

**Le volume des données** a des conséquences sur l'entreprise aux niveaux suivants :

Le changement dans le traitement de l'information a été déclenché aux États-Unis par les grandes entreprises de la distribution. Cela a conduit au lancement du processus appelé *Data Mining*.

- une plus grande capacité de stockage pour faire face à l'exponentiation des flux de données et d'informations nécessaires à la gestion des entreprises,
- une croissance extrêmement rapide de la quantité d'information entrant dans l'entreprise nécessitant des traitements de filtrage, de synthèse et de routage adéquat,
- une mise en place de nouveaux canaux de capture de l'information sous des formes moins traditionnelles (Internet, *Call Centers*, Télémarketing, *e-Business*, etc.) qui vont noyer l'entreprise non préparée et non flexible sous des quantités d'informations à tort inutilisées.

Cette révolution a d'abord touché aux Etats-Unis les grandes entreprises de la distribution qui voulaient exploiter les quantités astronomiques de données que fournissait chaque jour la prise en compte de l'information contenue dans les millions de tickets de caisses des clients achetants. L'exploitation systématique de cette manne incroyable de données a permis d'après des études qu'on a appelées Études *Data Mining* de trouver de façon simple les « causes à effets » de relations sous-jacentes dans le comportement d'achat des consommateurs. Ces relations de « séquences d'achats » et d'associations d'achats, une fois mises à jour, ont contribué à optimiser les processus d'approvisionnement et de stockage des produits vendus (le *Supply Chain Management* des Anglo-Saxons) ainsi que la logistique de placement des produits en magasin (la tendance étant alors de mettre côte à côte des produits dont l'achat de l'un entraîne l'achat de l'autre, maximisant dès lors les quantités achetées par les clients tout en minimisant les promotions globales). Cette approche fut un succès et fut déterminante dans le lancement du processus *Data Mining* (terme Anglo-Saxon difficilement traduisible qui signifie bien « fouille » des données) avec, à la clef, de façon sous-jacente, l'espoir de découvrir les « pépites » de ces mines de données, en d'autres termes des informations inconnues et discréminantes.

Le *Data Mining* s'est étendu à d'autres secteurs d'activités (banque, assurance, téléphonie…).

Les entreprises de la distribution furent rapidement suivies dans cette démarche par les entreprises du domaine bancaire (en particulier les banques à guichets), par les compagnies d'assurance, par les grands opérateurs de téléphonie (en particulier ceux de la téléphonie mobile) par des compagnies des eaux et d'électricité, plus récemment par les trans-

porteurs aériens ou ferroviaires et par certains industriels comme l'aéronautique, l'automobile, etc. et c'est bientôt tout le secteur économique américain qui a été touché par la vague du *Business Intelligence* dont le *Data Mining* est partie prenante.

Comme on le voit, c'est un domaine d'activité en pleine expansion, et derrière le grand « coup marketing » des Anglo-Saxons avec les termes de *Business lntelligence* et *Data Mining* (termes d'ailleurs difficilement traduisibles) il n'en reste pas moins vrai que ce domaine est concret, extrêmement prometteur, et difficile (d'où l'exigence de qualité et de compétence).

En France la vogue du *Business-lntelligence* et *Data Mining* a commencé à se répandre il y a quatre ans avec quelques trois ans de décalage par rapport aux États-Unis (ce qui est peu en comparaison de la téléphonie mobile, près de 7 ans ; ou Internet, près de 6 ans). Là, à l'inverse de ce qui s'est passé aux États-Unis, c'est dans la banque et dans l'assurance que ce processus a trouvé le meilleur écho, même si désormais ce sont les mêmes secteurs qu'aux États-Unis qui sont conquis et qui passent à l'action.

Le *Business Intelligence* n'est pas strictement parlant un nouveau domaine scientifique en soit. Il s'agit plutôt d'une industrialisation de méthodes statistiques et informatiques sur les bases de données.

Cette « vogue » n'est pourtant pas une « mode » comme l'ont été « l'intelligence artificielle » ou la « cognitique », inventées dans les années 70-80, mais bien une extension moderne, et sous une forme plus industrielle, de domaines scientifiques qui traitaient de l'analyse, de la prévision, du compactage, de l'estimation, calculés à partir des données ; ces domaines scientifiques portent un nom, il s'agit de la statistique au sens large : comprenant la statistique mathématique, les méthodes mathématiques de l'analyse des données et l'exploitation avancée des bases de données.

C'est donc bien d'industrialisation de méthodologies statistiques et informatiques sur les bases de données dont il est question et non d'un nouveau domaine scientifique en soi. Néanmoins, cette « industrialisation » va obliger le statisticien à apprendre de nombreuses matières, nouvelles pour lui, relatives aux domaines connexes dont il a été question précédemment. C'est ce que nous allons découvrir maintenant en entrant dans le détail des ajouts « méthodologiques » et « techniques » du *Business Intelligence* et du *Data Mining*.

## ▶ 3 – Les nouveaux paradigmes

Quelle est la modification essentielle dans le processus d'analyse et de traitement de l'information qui nourrit le débat aujourd'hui ? A-t-on attendu les Américains et leur approche du *Business lntelligence* pour prendre en compte les problèmes de gestion de la clientèle ou de l'optimisation des services ? Voici des questions que beaucoup se posent aujourd'hui face à l'invasion de termes qui, pour certains, restent

« barbares » et n'évoquent rien de précis sinon qu'ils imposent à l'entreprise des choix d'évolution.

La prise de conscience par les entreprises que le stockage systématique de données est coûteux, si l'on ne prend pas les moyens d'en exploiter la substantifique moelle, a commencé à donner le ton à cette tendance naturelle qui veut que l'on amoncelle avant de savoir à quoi cet amoncellement peut servir. Tout le mérite des Anglo-Saxons est d'avoir senti les premiers qu'un tournant se produisait visant à donner un « sens » et une « valeur » à l'information dont on dispose. Ce tournant se produisit dès lors que le paradigme latent consistant à considérer qu'une entreprise possède un capital fondamental : ses clients, ses produits et le capital intellectuel de ses cadres, s'est imposé de plus en plus clairement.

Les modifications essentielles dans le processus d'analyse et de traitement de l'information se sont produites lorsque les entreprises ont reconnu comme leur capital fondamental, leurs clients, leurs produits et le capital intellectuel de ses cadres.

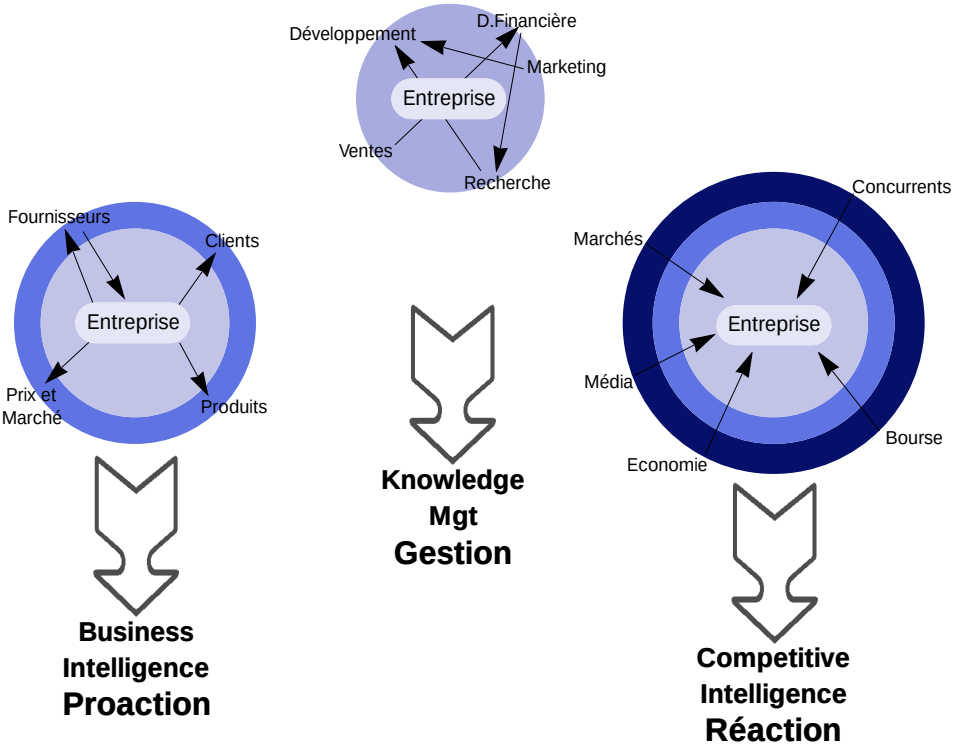
En effet, hier encore, une entreprise considérait que son capital était son outil de production, son bilan financier et les produits et services qu'elle délivrait au marché, ce dernier devant être conquis globalement. Aujourd'hui une entreprise se doit de travailler sur trois cercles concentriques qui touchent ses activités à des niveaux différents, toutes ses activités donnant lieu à analyse et remontées d'informations aux services décisionnaires, permettant ainsi une meilleure connaissance de son environnement. Ces trois cercles sont :

**1) Le cercle intérieur** : il s'agit là de l'ensemble des processus d'échange ou de transfert d'informations d'un service à l'autre, de mémorisation des « savoir-faire » ou des meilleures pratiques (*best practices*), de mise en commun des connaissances de l'entreprise en vue d'en optimiser la gestion ; les Anglo-Saxons ont donné un nom à ce cercle intérieur, ils l'appellent : *Knowledge Management* (gestion de la connaissance).

**2) Le cercle médian** : il s'agit de l'ensemble des techniques, des méthodes, des approches permettant à une entreprise d'agir directement en connaissance de cause sur ses clients, ses services et produits, ses fournisseurs, ses partenaires, ses processus, pour mieux servir son marché : ce premier cercle porte un nom, c'est le *Business Intelligence* dont il a déjà été question. Cette approche correspond à une attitude de « proaction » de la part de l'entreprise lorsqu'elle déploie les techniques associées.

**3) Le cercle externe :** il s'agit dans ce cas de l'ensemble des techniques, des méthodes, des approches mis à la disposition de l'entreprise lui permettant de mieux maîtriser son environnement extérieur lorsqu'elle n'a pas directement prise sur lui ; on trouvera ainsi dans ce cercle externe : les concurrents, les marchés, en général, la bourse, l'économie du domaine, les médias, etc. Cette approche du cercle externe correspond à une attitude de réaction à des pressions extérieures ; elle sera utile à l'entreprise lorsqu'elle veut prévoir son avenir et son environnement. Ce deuxième cercle porte un nom c'est l'intelligence économique (*Competitive lntelligence* des Anglo-Saxons).

**La figure ci-après donne les 3 cercles KM/BI/CI positionnés autour de l'entreprise**




A ce niveau, les statisticiens en place dans l'entreprise ou les cabinets d'études marketing auxquels les entreprises faisaient appel, se contentaient jusqu'ici de pratiquer des sondages sur la population des clients connus et d'effectuer à partir de ces échantillons extraits de la population mère, des « segmentations » permettant de déduire les principaux « profils » de clientèle pour lesquels l'entreprise devait satisfaire les attentes différenciées. Ici l'approche fusionnelle entre statistique et informatique va permettre d'analyser avec les outils fondamentaux du *Data Mining* la population mère exhaustivement, sans passer par l'utilisation des sondages, ce qui représente un changement drastique de comportement de l'analyste ou du statisticien. Le principe nouveau et fondamental sur lequel repose ce commentaire s'appelle principe de décomposition ; il est une caractéristique claire et nouvelle du changement à opérer dans les processus d'analyse, dès lors que l'on dispose d'outils pouvant supporter ce principe.

L'un des points essentiels dans l'approche de la *Business Intelligence* est la création par l'entreprise des structures de stockage des données codées dites *Data Warehouse* (entrepôt de données).

Le principe de décomposition stipule donc que pour garantir une homogénéité dans les populations à analyser dont dérivera une plus grande « robustesse » (au sens statistique du terme) des méthodes utilisées par la suite, qu'elles soient de *scoring* (quantification et probabi-

lisation des risques clients), de prédiction ou de recherche d'associations (associations d'achats de produits ou liens de cause à effet), il importe de décomposer la population initiale en regroupant dans les mêmes segments (classes) des individus « similaires », dès le départ, par des approches reposant sur l'exhaustivité.

Un *Datamart* (magasin de données) est une structure de stockage avec une forte orientation thématique.

Ce principe de décomposition est à la base du changement de l'approche analyse dont le nouveau métier de *Data Asset Manager* dans l'environnement Anglo-Saxon s'inspire, permettant ainsi de définir celui de « nouveau statisticien » dans une approche plus « française » d'analyse exploratoire des grandes bases de données qui en découlera.

En effet l'un des points essentiels, sous-jacent dans l'approche *Business Intelligence* aujourd'hui, est la création par l'entreprise des structures de stockage des données codées, dites de *Data Warehouse* (entrepôt de données). Le *Data Warehouse* est une collection de données orientées sujet, intégrées, non volatiles et historiées, organisées pour le support d'un processus d'aide à la décision. Les données sont dites orientées sujet car elles sont structurées par thèmes. Ceci permet au *Data Warehouse* d'être organisé autour des sujets majeurs de l'entreprise. Cette orientation sujet va également permettre le développement du système décisionnel au travers d'une approche incrémentale, sujet après sujet. L'intégration des différents sujets dans une structure unique est nécessaire pour ne pas avoir à dupliquer les informations communes à plusieurs sujets. Néanmoins, dans la pratique, cette orientation sujet (ou thème) peut être contenue dans une structure supplémentaire appelée *Datamart* (magasin de données). L'intégration des données s'effectue après une unification des données et leur validation qualitative.

La création d'un *Data Warehouse* nécessite donc un travail sur les données mais aussi un sérieux travail lié au choix technologique du modèle de données associé. Les méthodes de conception d'un modèle de données seront différentes selon le cadre d'utilisation du *Data Warehouse* ou du *Datamart* considérés.

Le *Data Mining* est un ensemble de techniques d'analyse des données spécialement adaptées pour traiter le grand volume de données stockées dans les *Data Warehouse*.

Dans le cadre décisionnel, le *Data Warehouse* est une base dédiée à la prise de décisions.

Le *Data Warehouse* dont il a été question précédemment est associé dans l'esprit de ceux qui parlent *Business lntelligence* aux nouveaux outils d'exploitation de ces structures de stockage que sont les outils et techniques de *Data Mining*. Le *Data Mining* est un ensemble de techniques d'analyse des données spécialement adaptées pour traiter les grands volumes de données stockées dans les *Data Warehouse*. Ils permettent d'extraire les informations les plus pertinentes, souvent non triviales, cachées au sein des données. En effet, lorsque l'on utilise des techniques d'analyse classique de type classification ou *clustering* sur de gros ensembles de données, il est nécessaire de travailler sur des échantillons représentatifs des populations étudiées. Les inconvénients sont alors nombreux.

En résumé, les facteurs nouveaux qui vont pousser au changement des métiers du statisticien ou de l'analyste d'entreprise (marketing, commercial, production) sont les suivants :

a) une prise de conscience par les entreprises que leur développement et leur croissance dépendaient de la connaissance totale du milieu extérieur ou intérieur : clients, concurrents, services, connaissances savoirs internes, etc. En d'autres termes, l'information était un « actif » (*asset*) essentiel au même titre que le capital et les outils de production. Actif qui bien que ne figurant pas dans les bilans comptables des entreprises aujourd'hui, s'y trouvera de facto (directement ou indirectement) dans un avenir proche.

b) Un besoin exprimé par le marché d'une part pour les techniques permettant une différenciation des offres produits ou services vis-à-vis de la concurrence (*Data Warehouse, Data Mining, Text Mining*) et l'exploitation analytique d'Internet (*Web-Mining* applicable dans le domaine du *e-Business* d'IBM), et d'autre part pour les outils de contrôle et de pilotage globaux (*War-room* ou *Management Cockpit*).

c) Une accélération de la diffusion d'outils puissants capables de traiter de grands ensembles de données, ceci étant dû à la fois à l'explosion des sources de données et d'information, et au principe de « copiage » de ce qui a réussi au voisin, très fréquent dans les marchés compétitifs.

d) Un nécessaire besoin de fusion de ces sources en des bases de données unifiées et uniques où données structurées et non structurées se mélangeront, donnant lieu à des processus d'analyse nouveaux et non encore réellement pratiqués aujourd'hui sur une grande échelle, donnant de ce fait un espace nouveau à l'investissement.

e) Un développement simultané des trois domaines nouveaux d'investissements des entreprises que sont : *Business lntelligence, Knowledge Management, Competitive lntelligence* auxquels il faut rajouter le domaine transverse et très en vogue aujourd'hui du *Customer Relationship Management*, domaine regroupant la mise en place des *call-centers*, les outils d'analyse déjà cités et les outils et techniques du « management de campagnes » (*campaign management* des Anglo-Saxons) donnant ainsi naissance au « Business de l'Information » réel.

*NB : Nous n'avons pas donné ici de bibliographie associée, celle-ci serait trop vaste et dépasserait de loin la taille du texte présenté ici, d'autre part elle serait pluridisciplinaire et très hétérogène regroupant des articles de type Marketing, Marketing Research, Économétrie, Business Management, Statistique appliquée, Statistique théorique, Mathématiques appliquées, Mathématiques pures, Linguistique, Linguistique computationnelle, Informatique des Bases de Données, Documentation et Bibliométrie, Scientométrie, etc.*