Livre Blanc :

La Révolution du Big Data…

Comment Extraire de la Valeur à partir des Big Data

Thomas HILL
Olivier LEBRET

Sommaire

Introduction

Le "Big data" est le terme à la mode que l'on retrouve actuellement dans toutes les conférences professionnelles en lien avec la data science, la modélisation prédictive, le data mining et le CRM, pour ne citer que quelques-uns des domaines littéralement électrisés par la perspective d'intégrer des jeux de données plus volumineux et des flux de données plus rapidement dans leurs processus métier et d'autres processus organisationnels. Comme c'est souvent le cas lorsque de nouvelles technologies commencent à transformer les industries, de nouvelles terminologies émergent, en même temps que de nouvelles approches pour conceptualiser la réalité, résoudre certains problèmes ou améliorer les processus.

Voilà encore quelques années, nous nous contentions de « segmenter » les clients en groupes susceptibles d'acquérir certains biens ou services spécifiques. Il est désormais possible et courant de construire des modèles pour chaque client en temps réel à mesure qu'il surfe sur Internet à la recherche de biens spécifiques : instantanément, les centres d'intérêt du prospect sont analysés et il est possible d'afficher des publicités ultra ciblées, ce qui constitue un niveau de personnalisation inconcevable il y a seulement quelques années.

Les technologies de géolocalisation des téléphones mobiles et de leur usagers sont matures, et la vision décrite dans le film de science-fiction *Minority Report* de 2002, où les publicités projetées dans l'enceinte des centres commerciaux ciblent directement les personnes qui passent devant, semble à portée de main. Inévitablement, la déception risque d'être à la hauteur des espérances dans de nombreux domaines tant les technologies autour du big data sont prometteuses. Un nombre restreint de données décrivant avec précision un aspect critique de la réalité (vital pour l'entreprise) est autrement plus précieux qu'un déluge de données relatives à des aspects moins essentiels de cette réalité.

L'objectif de cet article vise à clarifier et mettre en lumière certaines opportunités intéressantes autour du big data, et illustrer la manière dont les plates-formes analytiques ***STATISTICA*** de StatSoft peuvent exploiter cette profusion de données dans la perspective d'optimiser un processus, résoudre des problèmes, ou améliorer la connaissance client.

À Partir de Quand Peut-on Parler de Données Massives ou de Big Data ?

Bien évidemment, il n'existe pas de définition universelle, et la bonne réponse est "ça dépend". En fait, d'un point de vue pratique, et dans la plupart des discussions relatives à cette thématique, les big data se caractérisent par des jeux de données très volumineux, de l'ordre de plusieurs giga-octets à quelques téraoctets.

Ces données peuvent donc aisément être stockées et gérées dans des bases de données "traditionnelles" et avec du matériel informatique classique (serveurs de bases de données). Le logiciel ***STATISTICA*** est multitâches pour toutes les opérations fondamentales d'accès aux données (lecture), et pour tous ses algorithmes de transformation et de modélisation prédictive (et de scoring), ce qui permet d'analyser ces jeux de données (effectivement très volumineux) sans devoir utiliser de nouveaux outils spécialisés.

Gros Volumes de Données

Pour remettre les choses en perspective, certaines des plus grandes banques internationales, clientes de StatSoft, gèrent pour certaines entre 10 et 12 millions de comptes. Avec près de 1.000 paramètres ou caractéristiques (variables) par compte, organisés dans un entrepôt de données dédié au risque et aux autres activités de modélisation prédictive, ce type de fichier représente environ 100 giga-octets ; il ne s'agit pas de petits entrepôts de données, mais rien qui ne dépasse les capacités des technologies classiques des bases de données, et rien d'insurmontable pour ***STATISTICA***, même sur du matériel datant de plusieurs années.

En pratique, un très grand nombre d'applications d'aide à la décision dans le domaine de la santé, du secteur bancaire et financier, de l'assurance, de l'industrie manufacturière, etc…, s'appuient sur des bases de données souvent bien organisées de données clients, de données machines, etc... Dans la plupart des cas, la taille de ces bases de données, et la rapidité avec laquelle elles doivent être analysées pour répondre aux besoins métier essentiels de l'entreprise constituent de véritables défis. Les solutions de scoring et d'analyse par batch de ***STATISTICA*** (***STATISTICA Entreprise***), les solutions en temps réel (***STATISTICA Live Score***), ou les outils analytiques de création et de gestion de modèles (***STATISTICA Data Miner***, ***Plate-Forme Décisionnelle***), peuvent aisément être déployés sur plusieurs serveurs multiprocesseurs. Dans la pratique, les analyses prédictives (par exemple, de risque de crédit, de probabilité de fraude, de fiabilité des pièces produites, etc...) peuvent souvent être réalisées très rapidement pour les décisions nécessitant une réponse quasi-instantanée, grâce aux outils ***STATISTICA***, sans aucune personnalisation.

Gros Volumes de Données et Big Data

D'une manière générale, les discussions autour des big data se focalisent sur des entrepôts de données (et leur analyse) dépassant largement plusieurs téraoctets. Plus précisément, certains entrepôts de données peuvent dépasser plusieurs milliers de téraoctets, atteignant plusieurs pétaoctets (1.000 téraoctets = 1 pétaoctet). Au-delà des pétaoctets, les capacités de stockage des données se mesurent en exaoctets ; par exemple, le secteur de l'industrie manufacturière a stocké au total près de 2 exaoctets d'informations nouvelles en 2010 à l'échelle mondiale (Manyika et al., 2011).

Gros Volumes de Données, Big Data

Gros fichiers de données : de 1.000 mégaoctets (= 1 gigaoctet) à plusieurs centaines de gigaoctet

Très gros fichiers de données : de 1.000 giga-octets (= 1 téraoctet) à plusieurs téraoctets

Big Data : de plusieurs téraoctets à quelques centaines de téraoctets

Very Big Data : de 1.000 à 10.000 téraoctets (= 1 à 10 pétaoctets)

Dans certaines applications, les données s'accumulent très rapidement. Par exemple, pour les applications industrielles ou les chaînes de production automatisées, comme pour la production d'électricité, des flux de données continus sont générés chaque minute ou chaque seconde pour parfois plusieurs dizaines de milliers de paramètres. De la même

manière, nous avons vu apparaître au cours de ces dernières années la technologie du "smart-grid" pour des réseaux de distribution d'électricité « intelligents », qui permettent de mesurer la consommation électrique de chaque foyer minute par minute, voire seconde par seconde.

Pour ce type d'application, qui nécessite le stockage de données sur plusieurs années, il n'est pas rare de voir s'accumuler rapidement de très grosses volumétries de données (Hopkins et Evelson, 2011). Il existe de plus en plus d'applications dans l'administration et le secteur commercial où le volume de données et la vitesse à laquelle ces données sont accumulées nécessitent plusieurs centaines de téraoctets ou pétaoctets dédiés au stockage et à l'analyse des données. La technologie moderne permet aujourd'hui de suivre les individus et leur comportement de différentes manières, par exemple, lorsque nous surfons sur Internet, que nous achetons des produits sur Internet ou en grande surface (d'après Wikipédia, Walmart gère un entrepôt de données supérieur à 2 pétaoctets), ou que nous laissons notre téléphone portable activé en laissant des informations sur les endroits où nous sommes passés et où nous nous rendons. Les divers modes de communication, du simple appel téléphonique à l'information partagée sur les réseaux sociaux comme Facebook (30 milliards de post chaque mois selon Wikipédia), ou aux sites de partage vidéo comme You Tube (You Tube revendique l'envoi de 24 heures de nouvelles vidéos chaque minute ; source Wikipédia), qui génèrent des quantités massives de nouvelles données quotidiennes. De même, les technologies modernes de santé génèrent des quantités massives de données pour la délivrance de soins (images, films, suivi en temps réel) et le remboursement des organismes de santé.

Défis Techniques du Big Data

Il existe essentiellement trois types de défis autour du big data :

1. Le stockage et la gestion des données massives, de l'ordre de la centaine de téraoctets ou du pétaoctet, qui dépassent les limites courantes des bases de données relationnelles classiques du point de vue du stockage et de la gestion des données.
2. La gestion des données non-structurées (qui constituent souvent l'essentiel des données dans les scénarios big data), c'est-à-dire comment organiser du texte, des vidéos, des images, etc…
3. L'analyse de ces données massives, à la fois pour le reporting et la modélisation prédictive avancée, mais également pour le déploiement.

Stockage des Big Data

Les big data sont généralement stockées et organisées dans des systèmes de fichiers distribués. S'il existe différentes approches et stratégies de mise en œuvre, l'information est stockée sur plusieurs disques durs (parfois plusieurs milliers) et ordinateurs classiques. Un index (« map ») permet de savoir à quel endroit (sur quel ordinateur/disque) se situe une information particulière. En fait, dans un souci de sécurité et de robustesse, chaque information est généralement stockée plusieurs fois, par exemple sous forme de triplets.

Supposons par exemple que vous ayez collecté des transactions individuelles pour une enseigne de la grande distribution. Le détail de chaque transaction peut alors être stocké sous forme de triplets sur différents serveurs et disques durs, avec une table-maître ou « map » indiquant précisément l'endroit où est stocké le détail de chaque transaction.

Grâce à l'utilisation de matériel informatique classique et de logiciels libres pour la gestion de ce système de fichiers distribué (de type Hadoop), il est possible de créer assez facilement des entrepôts de données fiables de l'ordre du pétaoctet, et ces systèmes de stockage deviennent de plus en plus courants. Vous trouverez davantage d'informations techniques sur la mise en place de ces systèmes en recherchant sur Internet des informations relatives aux systèmes de fichiers distribués ou à Hadoop.

Informations Non-Structurées

La plupart des informations collectées dans les systèmes de fichiers distribués sont des informations non-structurées comme du texte, des images ou des vidéos. Ceci présente des avantages et des inconvénients. L'avantage, c'est que les entreprises et les administrations peuvent stocker « toutes les données » sans se préoccuper desquelles sont pertinentes et utiles dans une perspective décisionnelle. L'inconvénient, c'est qu'il est nécessaire de mettre en place des traitements massifs de données afin de pouvoir extraire des informations intéressantes. Si certaines de ces opérations peuvent être relativement simples (par exemple, calculer de simples effectifs, etc...), d'autres nécessitent des algorithmes plus complexes qui doivent être développés spécifiquement pour fonctionner efficacement sur le système de fichiers distribué (voir le point suivant).

Données et informations non-structurées. Comme ce fut le cas lors de la généralisation des bases de données relationnelles, le principal défi aujourd'hui, c'est que même si nous stockons de grandes quantités de données, seule l'information que nous pouvons en extraire les rend utiles. Par exemple, lors du déploiement de la plate-forme ***STATISTICA Entreprise*** sur le site de son entreprise industrielle, un cadre supérieur a déclaré à l'équipe StatSoft [qu'il avait] « dépensé une fortune en équipements informatiques et pour le stockage des données, sans en retirer le moindre bénéfice financier », car aucune réflexion n'avait été menée en amont sur la manière d'exploiter les données afin d'améliorer les activités au cœur même du métier de l'entreprise. En termes plus généraux, alors que la quantité de données croît de manière exponentielle, notre capacité à extraire de l'information et à agir sur cette information reste limitée et tend de façon asymptotique vers une limite (quelle que soit la manière dont les données sont stockées).

Il s'agit d'une considération essentielle que nous allons développer maintenant : les méthodes et procédures d'extraction et de mise à jour des modèles, dans une optique d'automatisation des décisions et des processus décisionnels, doivent être conçus en même temps que les systèmes de stockage des données afin de garantir l'intérêt et l'utilité de ces systèmes pour l'entreprise.

Analyse des Big Data

Il s'agit sans conteste du principal défi lorsque nous travaillons sur des données massives, souvent non-structurées : comment les analyser utilement. En fait, il existe beaucoup moins d'articles sur ce thème que sur celui des technologies et solutions de stockage permettant de gérer des big data. Il faut néanmoins tenir compte d'un certain nombre d'éléments.

Map-Reduce

En règle générale, lorsque nous analysons plusieurs centaines de téraoctets, voire plusieurs pétaoctets de données, il n'est pas réaliste d'extraire les données vers un autre endroit (par exemple, le serveur analytique ***STATISTICA Entreprise Server***) afin de les y analyser. Le

processus consistant à déplacer les données, au travers d'un câble, vers un serveur ou plusieurs serveurs distincts (pour un traitement parallèle) nécessiterait beaucoup trop de temps et de bande passante. En revanche, les calculs analytiques doivent être réalisés à proximité physique de l'endroit où sont stockées les données. Il est nettement plus simple d'amener l'analytique vers les données, que d'amener les données vers l'analytique. C'est exactement ce que permettent de faire les algorithmes map-reduce, c'est-à-dire des algorithmes analytiques conçus dans cette optique. Une composante centrale de l'algorithme va déléguer les sous-calculs en différents endroits du système de fichiers distribué puis combiner les résultats calculés par les nœuds individuels du système de fichiers (la phase de réduction). En résumé, pour calculer un effectif, l'algorithme va calculer, en parallèle dans le système de fichiers distribué, des sous-totaux dans chaque nœud, puis renvoyer à la composante maîtresse ces sous-totaux qui seront ensuite additionnés.

Il existe quantité d'informations sur Internet sur les différents calculs pouvant être réalisés à l'aide du patron d'architecture map-reduce, notamment pour la modélisation prédictive.

Statistiques Élémentaires, Business Intelligence (BI)

Pour les besoins BI de reporting, il existe différentes solutions open-source pour calculer des totaux, moyennes, proportions, etc… en utilisant map-reduce. Il est donc assez simple d'obtenir avec précision des effectifs et autres statistiques élémentaires pour le reporting.

Modélisation Prédictive, Statistiques Avancées

De prime abord, il peut sembler plus complexe de construire des modèles prédictifs en utilisant un système de fichiers distribué ; mais en pratique, ce n'est pas le cas pour différentes raisons.

Préparation des données. Rappelez-vous que l'essentiel des données que nous trouvons dans les systèmes de fichiers distribués pour du big data sont souvent des informations non-structurées (par exemple, du texte). En fait, il est plutôt rare de trouver des applications dans lesquelles les mesures ou les valeurs collectées génèrent des pétaoctets de données. Par exemple, StatSoft a conduit, avec d'excellents résultats, des projets complexes sur des jeux de données très volumineux, décrivant les opérations minute-par-minute dans des usines de production d'électricité, afin d'améliorer l'efficacité des équipements en place tout en réduisant les rejets polluants (Electric Power Research Institute, 2009). Si les jeux de données peuvent effectivement devenir très volumineux, comme pour toutes les données de processus en continu, l'information qu'ils contiennent peut être synthétisée. Par exemple, alors même que les données sont collectées seconde par seconde ou minute par minute, certains paramétrages spécifiques de clapets, d'arrivée d'air, de fourneaux et de température de gaz restent stables et invariants sur de longues durées. En d'autres termes, les données collectées seconde par seconde sont, pour l'essentiel, des réplications d'une même information. Par conséquent, une agrégation « intelligente » des données en amont (à l'endroit où sont stockées les données) est nécessaire et peut être réalisée simplement, ce qui permet d'obtenir des fichiers de données contenant toute l'information nécessaire sur les modifications dynamiques impactant l'efficacité et les rejets de gaz de l'usine en vue de la modélisation et de l'optimisation.

Analyse des ressentis clients et préparation des données. Cet exemple illustre bien que des jeux de données volumineux contiennent souvent des informations que nous pouvons synthétiser. Les données collectées par les compteurs électriques utilisant la technologie du

"smart-grid" vont posséder des caractéristiques similaires aux sentiments exprimés par une même personne sur un sujet particulier, ou même par un groupe d'individus sur un plus grand nombre de sujets. Ainsi, StatSoft a travaillé sur des projets de text mining concernant l'analyse de tweets relatifs la satisfaction des usagers par rapport à une compagnie aérienne spécifique et ses services. Sans grande surprise, si nous pouvons extraire un grand nombre de tweets pertinents sur une base horaire ou quotidienne, la complexité des sentiments qu'ils expriment est finalement assez limitée (et de faible dimension). La plupart des tweets sont des plaintes et des phrases courtes relatant de "mauvaises expériences." En outre, le nombre et l'intensité de ces sentiments est assez stable au cours du temps et pour des types de griefs spécifiques (par exemple, bagage perdu, vol annulé, nourriture de mauvaise qualité, etc…).

Par conséquent, dans cet exemple, une simple compression des tweets en scores de sentiments à l'aide de méthodes de text mining (telles que celles disponibles, par exemple, dans ***STATISTICA Text Miner*** ; voir Miner, Elder, Fast, Hill, Delen, Nisbet, 2012) permet d'obtenir des jeux de données de taille beaucoup plus faible qu'il est alors possible d'aligner plus facilement avec des données structurées existantes (système de billetterie, ou programme de fidélisation) afin de mieux appréhender la stratification des groupes de clients spécifiques et de leurs plaintes.

Il existe de nombreux outils pour réaliser ce type d'agrégation des données (par exemple, scoring de sentiments) dans les systèmes de fichiers distribués, et ce processus analytique peut donc aisément être mis en place.

Construction des Modèles

Nous devons parfois construire rapidement des modèles précis sur des big data stockées dans un système de fichiers distribué. En fait, il est généralement plus utile de construire un grand nombre de modèles sur des segments de données plus fins dans un système de fichiers distribué, mais nous reviendrons sur ce point ultérieurement.

En fait, vous pouvez déployer map-reduce pour différents algorithmes courants de data mining et de modélisation prédictive adaptés aux traitements massifs de données en parallèle pour des systèmes de fichiers distribués (et ces algorithmes peuvent être intégrés dans la plate-forme ***STATISTICA*** de StatSoft). Mais quand bien même vous pourriez accroître le volume des données de façon considérable, votre modèle de prédiction final s'en trouverait-il meilleur ou plus précis ?

Échantillonnage de Probabilités

En échantillonnage de probabilités, chaque observation de la population à partir de laquelle l'échantillon est tiré possède une probabilité connue d'être sélectionnée dans l'échantillon ; si la probabilité est identique pour chaque individu de la population, on parle de méthode de sélection à probabilité égale ou d'échantillon EPSEM (*equal probability of selection method* ; voir Kish, 1965, pour plus d'informations).

Comme indiqué dans un récent rapport de Forrester : "***Deux plus deux égal 3,9" est [souvent] suffisant*** (Hopkins & Evelson, 2011). La réalité statistique et mathématique suit cette logique : un modèle de régression linéaire utilisant, par exemple, 10 prédicteurs sur un échantillon de probabilité correctement tiré de 100.000 observations sera aussi précis qu'un modèle construit à partir de 100 millions d'observations.

Contrairement aux affirmations de certains acteurs dans la mouvance du big data qui prétendent que « toutes les données doivent être traitées », la vérité, c'est que la précision d'un modèle est fonction de la qualité de l'échantillon (chaque observation de la population devant avoir une probabilité connue d'être tirée) et de sa taille par rapport à la complexité du modèle. Peu importe la taille de la population !

C'est la raison pour laquelle, par exemple, on obtient généralement des résultats remarquablement précis dès la fermeture des bureaux de vote les soirs d'élection alors même que ces estimations ne reposent que sur quelques milliers de votants au plan national.

Échantillonnage map-reduce, compression des données, sélection des données. Quel impact sur l'analyse des big data ? Il existe différents algorithmes d'échantillonnage (map-reduce) très efficaces pour les systèmes de fichiers distribués. Ces algorithmes peuvent constituer une excellente approche pour exploiter ces données massives dans une perspective de modélisation prédictive simple et efficace, et obtenir rapidement un retour sur investissement vis-à-vis de l'infrastructure de stockage. Pour la plupart des applications concrètes, c'est une excellente stratégie, par exemple, en déployant les plates-formes ***STATISTICA Entreprise*** et ***Data Mining*** comme outil analytique en complément d'interfaces avec le système de fichiers distribué (les big data) qui réalisent les étapes de préparation des données/d'agrégation et/ou d'échantillonnage de probabilité à l'aide des algorithmes map-reduce (pilotés par la plate-forme ***Entreprise***).

Outre l'agrégation et l'échantillonnage des données, ce système peut également réaliser la nécessaire sélection détaillée des données (par exemple, sur la base d'une micro-segmentation de groupes spécifiques de clients) afin d'envoyer les données à la plate-forme analytique ***STATISTICA*** qui va alors construire des modèles précis pour des segments spécifiques (par exemple, des offres de services financiers destinées à des foyers à forte valeur).

Intégration de *STATISTICA* avec des logiciels libres. L'une des spécificités de la plate-forme ***STATISTICA Entreprise*** et ***Data Mining*** est qu'elle a été développée, dès le départ, comme une plate-forme de calcul pour l'entreprise, en utilisant des langages de programmation universels et des interfaces de données standard. Vous pouvez donc intégrer aisément dans cette plate-forme, outre les outils ultra-performants de StatSoft, des outils libres émergeants pour la gestion et la préparation des données, ou des procédures analytiques spécialisées utilisant la technologie map-reduce. Ces procédures sont gérées au travers de la plate-forme, comme n'importe quel autre nœud analytique dans les processus analytiques. Par exemple, la plate-forme *open-source* R est fréquemment utilisée pour mettre en œuvre des procédures et des calculs statistiques ultra-spécialisés, et la plate-forme ***STATISTICA*** est compatible avec la plate-forme R depuis de nombreuses années grâce à l'intégration élémentaire de scripts R dans des processus analytiques.

L'analyse et l'utilisation des big data est en plein émergence mais évolue aussi très vite. Il est important que la plate-forme analytique organisée autour du système de fichiers distribué puisse facilement intégrer de nouvelles méthodes de préparation et d'agrégation des

données, d'échantillonnage et de stratification afin de rentabiliser aussi rapidement que possible l'investissement réalisé dans le système de fichiers distribué.

Mise en œuvre de procédures spécialisées via map-reduce. Outre l'intégration aisée avec les plates-formes et autres outils *open-source*, il est important que la plate-forme analytique choisie offre la possibilité de personnaliser les processus analytiques afin de répondre aux besoins analytiques spécifiques basés sur le système de fichiers distribué et les big data. Les applications concrètes et les bonnes pratiques en matière d'analyse de données massives sont en pleine émergence et évoluent rapidement ; il n'existe pas de consensus universel contrairement à l'approche analytique et à l'analyse prédictive "traditionnelle" qui sont au contraire bien documentées. Toutefois, cette situation peut changer rapidement puisque l'ensemble des principaux éditeurs de bases de données et d'outils de BI (Microsoft, Oracle, Teradata et d'autres) offrent des interfaces et des outils permettant d'accéder et de traiter les données de façon efficace.

Dans tous les cas, la plate-forme ***STATISTICA Entreprise*** vous permet de construire votre propre implémentation des approches analytiques spécifiques utilisant les données des systèmes de fichiers distribués, mais reconnaît également les interfaces et outils clés-en-mains accessibles au travers des interfaces personnalisées des principaux éditeurs. Cette dernière approche constitue sans doute l'approche la plus efficace et la plus "naturelle" pour amener l'analytique vers les données massives (au travers de la plate-forme ***STATISTICA***).

Autres Considérations Pratiques pour la Mise en Œuvre

Pour synthétiser notre propos jusqu'à présent, les big data désignent essentiellement des informations non-structurées, stockées dans un système de fichiers distribué dont les données individuelles sont disséminées sur plusieurs centaines, voire plusieurs milliers de disques durs et de serveurs. La taille de ces systèmes de fichiers distribués peut aisément dépasser plusieurs pétaoctets (plusieurs milliers de téraoctets) avec les technologies actuelles. Pour les opérations élémentaires de préparation des données, de nettoyage, et d'extraction des données, il est plus efficace de réaliser les analyses respectives sur le site (le serveur spécifique) où sont stockées les données (afin de réduire et agréger les données sous forme de statistiques de synthèse) en utilisant une approche map-reduce.

Profiter de la Profusion de Données pour Construire un Grand Nombre de Modèles

Comme évoqué précédemment dans cet article, le véritable intérêt des données massives dans les systèmes de fichiers distribués n'est pas de calculer des modèles (prédictifs) globaux en utilisant l'intégralité des données disponible, mais plutôt des échantillons valides de données ; dans les deux cas, les résultats et la précision des modèles seront les mêmes.

En revanche, il est beaucoup plus pertinent d'utiliser cette profusion de données et les outils disponibles pour les segmenter efficacement, et construire un grand nombre de modèles avec de plus petites classes. Par exemple, nous pouvons nous attendre à ce que des modèles de montée en gamme (*upselling*) construits sur des segmentations plus larges (individus âgés de 20 à 30 ans) vont produire des résultats moins précis qu'un grand nombre de modèles construits sur des segmentations plus fines (des étudiants âgés de 20 à 21 ans, habitant en cité universitaire et bénéficiant d'un emploi salarié à temps partiel).

Par conséquent, une manière de tirer profit des big data et d'utiliser l'information disponible consiste à construire un grand nombre de modèles, avec un grand nombre de segments, et d'utiliser ces modèles pour scorer (prédire) les observations à l'aide du modèle le mieux adapté. En poussant à l'extrême, nous pourrions avoir un modèle distinct pour chaque individu d'un entrepôt de données gigantesque, afin de prédire ses futurs achats.

La plate-forme analytique connectée à l'entrepôt de données doit donc être en capacité de gérer plusieurs centaines ou plusieurs milliers de modèles, tout en offrant la possibilité de recalibrer ces modèles à volonté, si nécessaire. La ***Plate-Forme Décisionnelle de STATISTICA*** offre précisément ces fonctionnalités essentielles, et StatSoft jouit d'une grande expérience dans la construction de modèles automatisés et de calibrage en support de ces systèmes.

Déploiement des Modèles pour du Scoring en Temps Réel

L'une des composantes fondamentales de la plate-forme ***STATISTICA Entreprise*** est la solution ***STATISTICA Livescore®***, qui permet de scorer de nouvelles données en temps réel en utilisant les Services Web. Dans cet environnement, il est possible pour des programmes externes d'appeler des modèles gérés (avec contrôle de version) au travers de ***STATISTICA Entreprise*** afin de scorer de nouvelles données transmises soit directement au travers d'un appel distant au système, soit par l'intermédiaire d'un identifiant désignant une observation particulière ou un groupe spécifique d'observations à scorer.

En ce qui concerne les données massives et les systèmes de fichiers distribués, le processus de scoring est identique, qu'il s'agisse de données stockées dans des bases de données relationnelles ou de données stockées dans des systèmes de fichiers distribués. Le principal défi pour maintenir des performances acceptables réside au niveau de la gestion et de la préparation des données, mais ces étapes peuvent être réalisées à l'aide des outils map-reduce disponibles pour la préparation et l'extraction des données ou en envisageant d'autres architectures. Il peut s'agir d'entrepôts de données spécifiques, dédiés à l'analytique et au scoring, basés sur des bases de données relationnelles et alimentés par des routines ETL map-reduce, ou l'une des technologies émergentes basés sur les RAM clouds, où le système de fichiers distribué est stocké lui-même sur des « disques » mémoire très rapides pour des temps d'accès très rapides aux données. Il existe également un certain nombre de solutions commerciales disponibles, comme Netezza, Oracle Extreme, SAP Hanna, etc..., qui ont toutes été conçues dans un même objectif : accéder à des entrepôts de données gigantesques de façon très rapide. Bien entendu, ***STATISTICA*** peut s'intègrer avec l'ensemble de ces solutions.

Critiques des Stratégies Big Data, Stratégies de Mise en Œuvre

StatSoft existe depuis une trentaine d'années et conseille ses clients en matière de meilleures pratiques analytiques afin de leur garantir un retour sur investissement rapide. StatSoft se limite ***uniquement*** à son rôle d'éditeur de solutions analytiques, et ne commercialise ni matériel, ni solution dédiée de stockage. Au cours de ces années, nous avons connu diverses technologies nouvelles qui sont passé par le cycle habituel d'un grand engouement initial, du succès pour les pionniers, puis de maturation au travers de solutions et processus standard afin d'optimiser le ROI.

Inévitablement dans ce schéma, il peut y avoir des déconvenues et des désillusions lorsque les promesses initiales des nouvelles technologies ne se matérialisent pas. Il faut donc conserver un certain nombre d'éléments à l'esprit :

Les Données Massives n'apportent pas Nécessairement une Meilleure Connaissance (Client)

Supposons que vous ayez accès au cours de toutes les actions de la bourse de Paris, et que vous ayez enregistré ces cours, seconde par seconde, avec au final une masse de données pharaonique. Imaginez par ailleurs que vous ayez accès à un certain nombre d'informations sensibles et ciblées concernant certains indicateurs et résultats financiers de la société. Quelle stratégie faut-il privilégier pour devenir un bon trader ? Sans nul doute, la seconde !!!

Le fait de stocker des quantités massives de données décrivant des phénomènes aisément observables de la réalité ne se traduit pas nécessairement par une meilleure connaissance de cette réalité. C'est la même chose si vous analysez des cours boursiers, des tweets, des données médicales ou marketing, ou des données de machineries complexes afin de réaliser de la maintenance prédictive.

Par exemple, il peut être beaucoup plus pertinent pour un magasin d'ameublement et de décoration de disposer d'une liste fiable de prospects intéressés par des meubles et accessoires de maison, avec des informations démographiques et des indicateurs des revenus du ménage, plutôt qu'une quantité massive de données sur le parcours de navigation en ligne sur différents sites d'ameublement. Pour la performance d'une usine de production d'électricité, nous avons montré [voir Electric Power Research Institute (EPRI), 2009] qu'il est important de se focaliser sur certaines informations particulières et sur les modifications intervenant sur certaines combinaisons de paramètres pour prédire l'évolution des performances et des rejets futurs, plutôt que de suivre plusieurs milliers de paramètres, seconde par seconde.

Comme c'est le cas avec tout projet visant à optimiser les performances organisationnelles et métier, il est important de se poser un certain nombre de questions du type : « À quoi devraient ressembler des résultats idéaux », « Comment puis-je mesurer la réussite de mon projet (savoir quand j'ai terminé, et ce que j'ai gagné) » ou « De quelles informations vais-je avoir besoin pour tendre vers les résultats idéaux ». Les réponses à ces questions peuvent parfaitement conduire à la mise en place d'un entrepôt big data avec une plate-forme analytique ; mais bien souvent, ce ne sera pas le cas.

Vélocité des Données et Temps de Réaction

Un autre aspect à prendre en compte concerne la vélocité des données ou la rapidité avec laquelle les données se mettent à jour. La vraie question concerne le « temps de réaction » nécessaire. En d'autres termes, vous pouvez construire des modèles dans un environnement de production pour prédire des problèmes imminents, avec une seconde d'avance, sur la base de données collectées en continu sur des milliers de paramètres. Toutefois, si un ingénieur a besoin de deux heures pour comprendre et prendre les mesures correctives nécessaires, ce système est absolument inutile.

De la même manière pour un magasin d'ameublement, il est plus important de recevoir une « alerte » un ou deux mois ***avant*** une transaction immobilière, plutôt qu'une information en temps réel après conclusion de cette transaction, lorsque le prospect a déjà commencé à prospecter sur Internet pour acheter ses meubles. Une alerte précoce permettrait à un

professionnel d'engager diverses démarches auprès de ce prospect ***avant*** qu'il n'entame son processus d'achat de meubles, pour lui proposer des offres spéciales ou le convaincre de se rendre dans un magasin et construire une relation personnelle privilégiée avec l'enseigne. À nouveau, une plate-forme en temps réel du parcours de navigation sur Internet peut ne pas être l'entrepôt de données idéal pour générer du trafic et construire une base clients loyale.

D'une manière générale, la bonne approche consiste à définir soigneusement, dès le départ, quel sera l'usage final ainsi que les stratégies pour réussir. À ce stade, les délais de réaction nécessaires (« à quel moment l'alerte doit-elle être déclenchée ») deviennent évidents, ce qui permet de définir le besoin, d'où découle évidemment un système optimal de collecte et de stockage des données, ainsi qu'une stratégie analytique.

Synthèse

L'objectif de cet article vise à dresser un bref aperçu des défis spécifiques posés par les big data, c'est-à-dire des entrepôts de données de l'ordre du téraoctet jusqu'à plusieurs pétaoctets de données (voire davantage), et des technologies et approches permettant de relever ces défis afin d'extraire de la valeur de ces données massives.

La technologie des systèmes de fichiers distribués, déployée sur des serveurs et systèmes de stockage courants, a rendu possible mais aussi économiquement viable, la création et la maintenance de ces entrepôts. Sur ces systèmes, au lieu de stocker les données sur un seul système de fichiers, les données sont stockées et indexées sur plusieurs (parfois des milliers de) disques durs et de serveurs, avec une indexation (la partie ***map***) redondante permettant de connaître l'endroit spécifique où certaines informations spécifiques se situent. Hadoop est sans doute le système le plus répandu à ce jour, qui utilise cette approche.

Pour traiter des données dans un système de fichiers distribué, il est nécessaire que les calculs simples tels que la comptabilisation des effectifs, la préparation et l'agrégation élémentaire des données, etc..., s'effectue à l'emplacement physique où se situent les données dans le système de fichiers distribué, plutôt que de déplacer les données vers le moteur de calcul analytique. La partie ***map*** des algorithmes de calcul respectifs contrôle ensuite les résultats individuels puis les agrège (la partie ***reduce***) ; ce schéma pour la mise en œuvre des algorithmes de calcul est connu sous le nom ***Map-Reduce***.

En pratique, la véritable valeur des big data se situe rarement dans le calcul des résultats statistiques réalisés sur l'***intégralité*** des données ; en fait, il existe des fondements statistiques qui montrent que ces calculs ne produisent de résultats plus précis dans la plupart des cas. En revanche, la véritable valeur des big data, en particulier pour du data mining ou de la modélisation prédictive, réside dans notre capacité à "micro-segmenter" l'information disponible en petits groupes, et à construire un grand nombre de modèles spécifiques pour ces petits groupes d'observations. Nous avons également abordé d'autres considérations généralistes concernant la valeur des big data dans cet article.

Du point de vue de la mise en œuvre, la plate-forme d'analyse des big data doit être en mesure d'intégrer les technologies émergentes des algorithmes, qui sont souvent des projets libres du domaine public. La ***Plate-Forme Décisionnelle de STATISTICA*** offre toutes les fonctionnalités pour exploiter les données massives, et gérer plusieurs milliers de modèles sur ce type de données.

Références

Eaton, C., Deroos, D., Deutsch, T., Lapis, G., Zikopoulos, P. (2012). ***Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming data.*** McGraw-Hill.

Economist Intelligence Unit Limited (2011). ***Big data: Harnessing a game-changing asset.*** The Economist.

Electric Power Research Institute (EPRI) / StatSoft Project : ***Statistical Use of Existing DCS Data for Process Optimization***; Palo Alto, 2009 (principal contributeur : Thomas Hill, StatSoft Inc. ; voir aussi http://my.epri.com/portal/server.pt?Abstract_id=000000000001016494).

Hopkins, B., & Evelson, B. (2011). ***Expand your digital horizon with Big Data.*** Forrester Research Inc.

Kish, L. (1965). ***Survey sampling.*** New York: Wiley.

Manyika, J., Chi, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). ***Big data: The next frontier for innovation, competition, and productivity.*** McKinsey Global Institute.

Miner, G., Elder, D., Fast, A., Hill, T., Delen, D., Nisbet, R. (2012). ***Practical text mining and statistical analysis for non-structured text data applications.*** Academic Press.

Glossaire

Big Data

Généralement, les discussions sur le thème des big data dans le cadre de la modélisation prédictive et du data mining concernent des entrepôts de données (et les analyses réalisées sur ces entrepôts de données) dépassant plusieurs téraoctets de données (1 téraoctet = 1.000 giga-octets ; 1 gigaoctet = 1.000 mégaoctets). Certains entrepôts de données peuvent même atteindre plusieurs milliers de téraoctets, c'est-à-dire de l'ordre du pétaoctet (1.000 téraoctets = 1 pétaoctet). Au-delà du pétaoctet, le stockage des données se mesure en exaoctets ; par exemple, on estime que l'industrie manufacturière, à stocké, tout secteur confondu, environ 2 exaoctets de nouvelles information à l'échelle mondiale en 2010 (Manyika et al., 2011).

Système de Fichiers Distribué

Les big data (plusieurs téraoctets, pétaoctets) sont généralement stockées et organisées dans des systèmes de fichiers distribués. S'il existe différentes approches et stratégies de mise en œuvre, l'information est stockée sur plusieurs disques durs (parfois plusieurs milliers) et ordinateurs classiques. Un index (« map ») permet de savoir à quel endroit (sur quel ordinateur/disque) se situe une information particulière. En fait, dans un souci de sécurité et de robustesse, chaque information est généralement stockée plusieurs fois, par exemple sous forme de triplets.

Supposons par exemple que vous ayez collecté des transactions individuelles pour une enseigne de la grande distribution. Le détail de chaque transaction peut alors être stocké sous forme de triplets sur différents serveurs et disques durs, avec une table-maître ou « map » indiquant précisément l'endroit où est stocké le détail de chaque transaction.

Grâce à l'utilisation de matériel informatique classique et de logiciels libres pour la gestion de ce système de fichiers distribué (de type Hadoop), il est possible de créer assez facilement

des entrepôts de données fiables de l'ordre du pétaoctet, et ces systèmes de stockage deviennent de plus en plus courants.

Exaoctet

1 exaoctet équivaut à 1.000 pétaoctets, soit 1.000 * 1.000 téraoctets.

Hadoop

Il s'agit d'un système de fichiers distribué permettant de stocker et gérer des entrepôts de données allant de plusieurs téraoctets à quelques pétaoctets.

Map-Reduce

En règle générale, lorsque nous analysons plusieurs centaines de téraoctets, voire plusieurs pétaoctets de données, il n'est pas réaliste d'extraire les données vers un autre endroit afin de les y analyser. Le processus consistant à déplacer les données, au travers d'un câble, vers un serveur ou plusieurs serveurs distincts (pour un traitement parallèle) nécessiterait beaucoup trop de temps et de bande passante. En revanche, les calculs analytiques doivent être réalisés à proximité physique de l'endroit où sont stockées les données. Il est nettement plus simple d'amener l'analytique vers les données, que d'amener les données vers l'analytique.

C'est exactement ce que permettent de faire les algorithmes map-reduce, c'est-à-dire des algorithmes analytiques conçus dans cette optique. Une composante centrale de l'algorithme va déléguer les sous-calculs en différents endroits du système de fichiers distribué puis combiner les résultats calculés par les nœuds individuels du système de fichiers (la phase de réduction). En résumé, pour calculer un effectif, l'algorithme va calculer, en parallèle dans le système de fichiers distribué, des sous-totaux dans chaque nœud, puis renvoyer à la composante maîtresse ces sous-totaux qui seront ensuite additionnés.

Pétaoctet

1 pétaoctet = 1.000 téraoctets.

Téraoctet

1 téraoctet = 1.000 giga-octets. La technologie actuelle des systèmes de fichiers distribués comme Hadoop permet de stocker et gérer plusieurs téraoctets de données dans un même entrepôt.