Le Big Data

Le terme même de Big Data a été évoqué la première fois par le cabinet d'études Gartner en 2008 mais des traces de la genèse de ce terme remontent à 2001 et ont été évoquées par le cabinet Meta Group.

Il fait référence à l'explosion du volume des données (de par leur nombre, la vitesse à laquelle elles sont produites et leur variété) et aux nouvelles solutions proposées pour gérer cette volumétrie tant dans la capacité à stocker et explorer celles-ci que, récemment, la capacité à analyser et exploiter ces données dans une approche temps réel.

Pourquoi le Big Data devient un élément clé des SI ?

> Les données non structurées, un vivier d'informations inexploitées

« Il n'y aurait que 10% de données structurées en entreprise, la quasi-totalité constituant ce remarquable bazar ambiant qui va de la messagerie électronique, en passant par les .pdf, les .ppt et autres joyeuses abréviations désignant aussi bien des fichiers texte, qu'audio ou vidéo.

Le problème, c'est que ces données ont la fâcheuse tendance à doubler en volume tous les deux mois, ce qui représente la bagatelle d'une croissance de +6.400% l'an (source IDC) ! A l'inverse, la croissance des données structurées ne connaîtrait qu'une (petite) croissance de 4% par an (source OTC). » © OpenText

Outre la production de documents internes, une des sources des données non structurées réside dans les échanges client par l'intermédiaire d'e-mails, mais aussi par la combinaison de technologies, comme les courriers/dossiers papier (reconnaissance de texte) et les échanges téléphoniques (vocal vers fichier texte).

> Un chiffre clé, l'accroissement du volume de données

« Chaque jour, nous générons 2,5 trillions d'octets de données. A tel point que 90% des données dans le monde ont été créées au cours des deux dernières années seulement.

Ces données proviennent de partout : de capteurs utilisés pour collecter les informations climatiques, de messages sur les sites de médias sociaux, d'images numériques et de vidéos publiées en ligne, d'enregistrements transactionnels d'achats en ligne et de signaux GPS de téléphones mobiles, pour ne citer que quelques sources. » © IBM

Cet accroissement de volume est principalement lié, dans les secteurs de la banque, de l'assurance ou encore des opérateurs, à la volonté de ces derniers de sans cesse mieux connaître leurs clients en croisant l'ensemble des informations disponibles sur celui-ci et sur ces actions quelle qu'en soit l'origine.

Derrière le Big Data, un concept clé : les 5 Vs

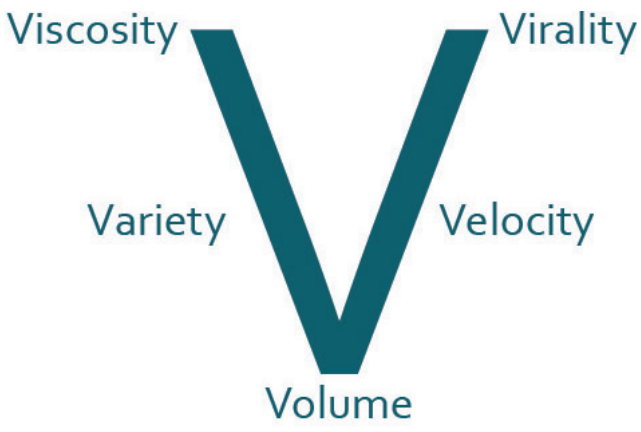
Volume : représente la quantité de données générées, stockées et exploitées au sein du SI. L'accroissement au sein des SI de la volumétrie s'explique par l'augmentation de la quantité de données générées et stockées mais aussi et surtout par le besoin d'exploiter des données qui, jusqu'à présent, ne l'étaient pas.

Variety : représente la démultiplication des types de données gérées par un SI, on parle ici de type de données au sens fonctionnel du terme et non pas uniquement au sens technique. Démultiplication qui entraîne aussi la complexification des liens et des types de lien entre ces données.

Velocity : représente la fréquence à laquelle les données sont générées, capturées et partagées. Les données arrivent par flux et doivent être analysées en temps réel pour répondre aux besoins des processus chrono-sensibles.

Viscosity : représente la résistance à laquelle se heurte l'organisation pour explorer, exploiter les données disponibles au sein des processus métier.

Virality : représente la capacité à diffuser rapidement l'information dans l'organisation afin de permettre la prise en compte de celles-ci au sein des processus métier.



Derrière le concept, des approches

Comme nous l'avons vu, le Big Data réside dans la capacité à gérer en temps réel un volume sans cesse croissant et changeant de données issues de diverses sources. Afin d'approfondir les solutions répondant à ce besoin, il convient de distinguer les différents cas d'usage que l'on va décliner en fonction du type de données manipulées et de l'usage que l'on désire faire de ces données.

Données non structurées, le virage du Text Mining

Le Text Mining (fouille de textes) permet au sein d'un ensemble de documents d'effectuer une analyse de leur contenu au travers d'une recherche sémantique reposant sur l'analyse du langage naturel (le français par exemple) et la gestion d'ontologies[1] spécialisées (pour un secteur d'activité, un métier). Cette fouille peut permettre de déterminer le contenu d'un document, mais aussi d'aller jusqu'à faire de l'analyse de sentiment au travers des tournures de phrases afin de savoir par exemple si un client se plaint ou fait une simple demande d'information.

A l'issue de cette fouille, on produit la liste des « concepts et relations » [2] abordés dans un document afin de pouvoir alimenter une base de connaissances qui permet :

- soit d'effectuer des recherches au sein de ce fond documentaire,
- soit d'extraire des données qui serviront à alimenter d'autres systèmes.

La différence entre une analyse sémantique et une indexation classique de document est que l'indexation se contente de référencer les mots présents dans un document sans s'intéresser au sens, à l'usage fait de celui-ci.

Les données non structurées

Elles sont définies, par opposition, comme des données disponibles mais non directement exploitables. De fait il s'agit des données que l'on peut extraire de tous types de documents électroniques (e-mail, document Word, vidéo, image, SMS, courrier digitalisé, page Web, réseau social).

Données structurées, le virage de la Big Analytic (ou Big Data Analytic)

Dans cette approche, l'analyse des données structurées évolue de par la variété et la vélocité des données manipulées. On ne peut donc plus se contenter d'analyser des données et de produire des rapports, la grande variété des données fait que les systèmes en place doivent être capables d'aider à l'analyse des données.

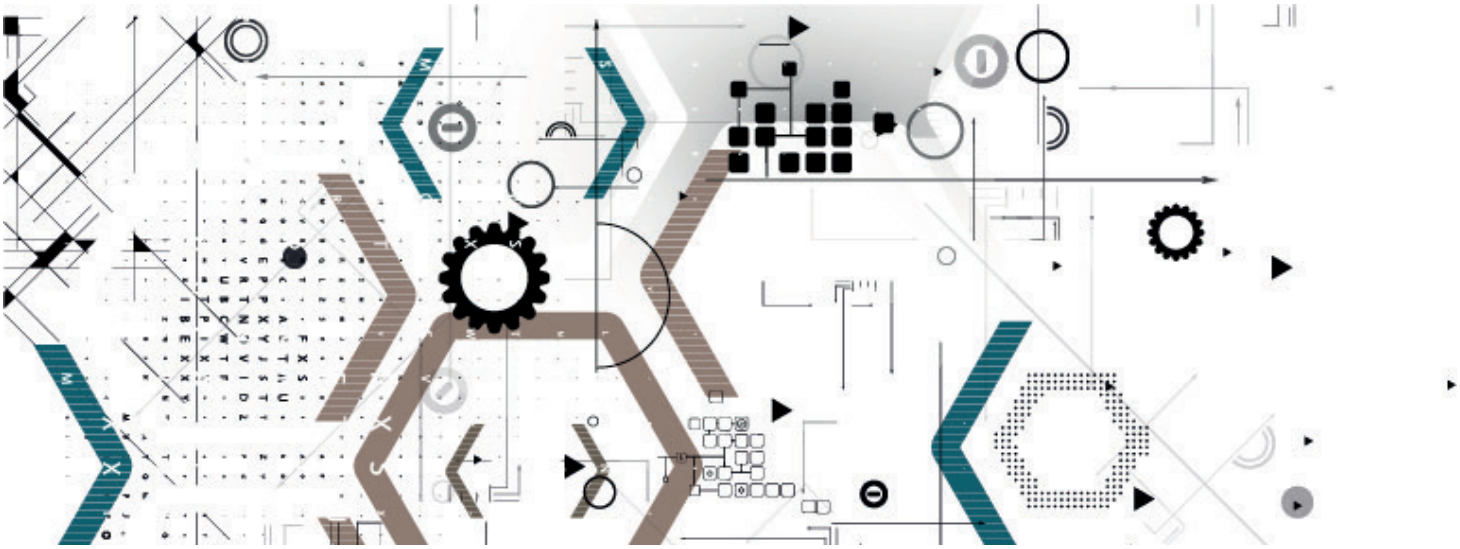
L'analyse consiste à déterminer, de façon automatique, au sein d'une variété de données évoluant rapidement les corrélations entre les données afin d'aider à l'exploitation de celles-ci.

Les données structurées

Elles sont définies par le fait qu'elles sont disposées de façon à être traitées automatiquement et efficacement par un logiciel, mais non nécessairement par un humain.

Le Text Mining et la Big Analytic, un lien possible

La Big Analytic repose sur l'analyse de données structurées. Mais comme, nous l'avons vu les données non structurées constituent un vivier d'informations qui devient essentiel. D'où la question : comment exploiter ces données dans la cadre de la Big Analytic ? Une des solutions consiste à utiliser le Text Mining pour retrouver des données clés dans des éléments non structurés afin d'alimenter un référentiel de données structurées que pourra exploiter la Big Analytic.



[1] On appelle « ontologie » un ensemble structuré de termes et de concepts représentant le sens d'un champ d'informations. Appliquée au Text Miming il s'agit donc d'un modèle de données conceptualisé sous forme de graphe qui définit l'ensemble des concepts liés à un domaine et la façon dont sont liés ces concepts (cf. notion de concept et relation).

[2] La notion de « concept et relation » est issue de la sémantique sur laquelle reposent les solutions des Text Mining. Elle définit l'extraction de mots-clés issus d'un langage courant ou spécifique à un métier (les concepts) et de liens entre ces mots-clés déterminés à partir de la structure grammaticale d'une phrase ou d'un paragraphe (les relations).

Le Big Data et ses technologies

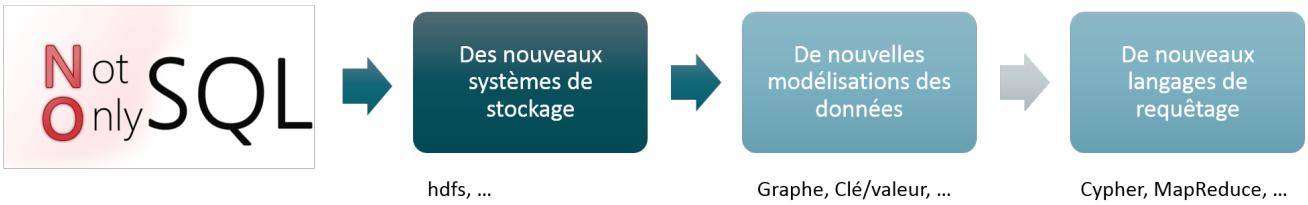
Si aujourd'hui le Big Data est possible, c'est grâce aux évolutions technologiques (logicielle ou hardware) qui permettent de répondre au 5Vs et aux usages nouveaux que l'on souhaite faire des données.

> Les solutions de stockage

Les Big Database : il s'agit de nouvelles solutions de SGBD permettant de gérer de gros volumes de données, dans une approche de variété et de vélocité fortes. Ces solutions reposent sur de nouveaux systèmes de gestion de fichiers partagés et permettent de gérer différemment le stockage, la modélisation et le requêtage des données.

On retrouve derrière ces nouvelles solutions le concept de SGBD NoSQL.

D'abord défini comme « NO SQL », ce terme a très vite évolué vers « Not Only SQL ».Ce concept exprime la possibilité de sortir d'une approche classique relationnelle (SGBD/R) sur le stockage et la manipulation des données.



Wide Column Store

Avec une approche de stockage des informations par colonne et non par ligne, comme les SGBD relationnel classique. Ces solutions offrent comme avantage une meilleure capacité d'évolution de la structure d'une table et, pour le stockage, la capacité de compresser les données de façon plus efficace.

Les principaux acteurs : Cassandra, BigTable, Hive, Hbase, Cloudera.

Key Value / Tuple Store

Ces solutions permettent de gérer les informations sous forme de couple clé/valeur liés entre eux pour former un enregistrement (row). Elle offre l'avantage de permettre de créer des enregistrements variables dans les données qui les constituent, contrairement à une approche de type table (base relationnelle) où les colonnes définissant un enregistrement sont fixes.

Les principaux acteurs : Amazon Dynamo, LevelDB, FundationDB, BerkeleyDB, Memcache DB. Caché (InterSystem), PI (OsiSoft).

Document Store

Ces solutions permettent de stocker et de gérer des documents. Elles ont une approche de gestion de données semi-structurées permettant de définir et d'associer des métadonnées à un document et de gérer la classification de ceux-ci.

Les principaux acteurs : MongoDB, CouchDB, RethinkDB, TerraStore, SimpleDB, Riak.

Graph

Ces solutions sont conçues pour optimiser la gestion de relations (notion d'arc en théorie des graphes) entre des objets (notion de nœud). L'idée principale est de permettre de retrouver des informations par les liens qui les unissent. Ces systèmes sont donc particulièrement adaptés aux moteurs sémantiques car ils permettent de modéliser facilement les concepts et relations qui sont le cœur de la sémantique.

Les principaux acteurs : Neo4j, InfiniteGraph, HyperGraphDB, InfoGrid, Trinity.

Bien sûr, les solutions de SGBD plus « classiques » restent utilisables de par l'évolution de leur architecture. C'est le cas pour les bases relationnelles, les bases objets, les bases multidimensionnelles, dont les capacités ont évolué au travers des architectures massivement parallèles ou InMemory[3].

[3] Le terme « InMemory » désigne les nouvelles architectures de gestion de bases de données (IMDB) qui utilisent la mémoire vive des serveurs pour le stockage des données, permettant ainsi un accès plus rapide à celles-ci. De plus, les IMDB gèrent la répartition des données sur plusieurs serveurs et leur réplication sur des supports de stockage physique (HDD, SDD, etc.) afin de garantir le support ACID (Atomicity, Consistency, Isolation, Durability) indispensable à un SGBD.

> Les nouvelles solutions logicielles

Les moteurs sémantiques (Text Mining) : généralement couplés avec un moteur de recherche, ils permettent de faire une analyse sémantique des documents afin d'en comprendre le contenu et ainsi de permettre de retrouver, au sein d'une base documentaire, le(s) document(s) traitant d'un sujet, parlant d'une personne.

Parmi les solutions les plus connues : Fise, Zemanta, iKnow (InterSystems), Noopsis, Luxid (Temis), LingWay.

Les solutions d'analytique : ce sont des solutions qui permettent de gérer la variété des données exploitées par une visualisation nouvelle de celles-ci avec une première analyse qui les contextualise, compartimente, corrèle. Pour cela, ces nouvelles solutions cherchent à aller au-delà d'une analyse statistique des données pour aller vers une analyse prédictive et la prise en compte de la temporalité des données.

Parmi les solutions les plus connues : QlickView, PowerPivot, Tableau.

Ainsi que, pour la manipulation des données : Aster, Datameer, SPSS, SAS ou Kxen pour le DataMining.

> Les matériels et les architectures

La puissance de calcul : inutile de revenir sur la fameuse loi de Moore qui prédit un doublement annuel de la puissance des processeurs. A celle-ci, s'ajoute les capacités d'algorithmes de type MapReduce, du Grid Computing ou des architectures massivement parallèles de type Appliances qui offrent à moindre coût l'équivalent de supercalculateurs, à l'instar de ce que propose Oracle avec Exadata ou IBM avec Netezza ou BCU.

Les capacités de stockage : l'évolution du stockage vers des systèmes distribués où un même fichier peut être réparti sur plusieurs systèmes permet d'envisager des volumes de stockage qui étaient auparavant inconcevables. Les technologies même de stockage évoluent pour offrir des accès toujours plus rapides à la donnée.

Le cloud : la capacité de stockage et la puissance de calcul devient un consommable de base au même titre que l'eau ou l'électricité. Vu sous l'angle « Big Data », ceci ouvre de nouveaux horizons, puisqu'au lieu de dimensionner les infrastructures pour les pics de stockage ou de traitement, les organisations peuvent désormais ajuster la taille et donc le coût de leurs infrastructures de calcul et de stockage au gré de l'évolution de leurs besoins.

Vous voulez en savoir plus

Sur le Big Data en général : wikipedia.org/wiki/Big_data ou www.techrepublic.com/blog/big-data-analytics.

Sur les technologies et les approches :

- Le NoSQL et les nouvelles solutions de SGBD : nosql-database.org
- La sémantique, ce que cela apporte et comment : www.proxem.com/2012/04/25/recherche-semantique-33-quap-porte-lanalyse-semantique

Vous voulez approfondir avec nous ?

Dans le cadre des recherches de la cellule Innovation, Aubay va poursuivre sur la fin de l'année 2013 plusieurs projets de veille technologique afin d'approfondir son expertise technique autour du Big Data et plus particulièrement des SGBD NoSQL et du Text Mining.

Si vous êtes intéressé soit pour être tenu au courant de l'avancée de ces travaux et des résultats obtenus soit pour participer avec nous au déroulement de ces travaux, n'hésitez pas à nous contacter (innov-dt@aubay.com).d'un plug-in particulier ou d'un device spécifique (type LeapMotion ou Kinect).

Vous souhaitez rejoindre une des premières ESN européennes ?

Véritable aventure entrepreneuriale, Aubay est la plus importante Entreprise de Services du Numérique née après 1998. 9ème ESN cotée sur EURONEXT PARIS, le Groupe Aubay compte en 2014 plus de 3 300 collaborateurs en Europe et près de 2 000 collaborateurs en France.

aubay TV www.youtube.com/AubayTV

@groupeaubay

www.aubay.com • nos offres d'emploi et de stage