TECHNIQUES
DE L'INGÉNIEUR

Réf. : **H6040 V1**

Date de publication :
10 février 2017

Introduction au Big Data - Opportunités, stockage et analyse des mégadonnées

Cet article est issu de : **Technologies de l'information | Technologies logicielles Architectures des systèmes**

par **Bernard ESPINASSE, Patrice BELLOT**

Mots-clés
big data | Stockage | mégadonnées | analytique

Résumé L'objet de cet article est de cerner ce terme Big Data ou mégadonnées. Dans un premier temps, les mégadonnées sont caractérisées au travers du modèle des 3V étendu au 5V. La problématique des mégadonnées est distinguée de celle de l'informatique décisionnelle. Les enjeux économiques et sociétaux associés aux mégadonnées sont abordés en présentant différents exemples d'usage relevant de différents domaines d'activité. Sont ensuite introduites différentes grandes méthodes et techniques associées au stockage et à l'exploitation/analyse de ces mégadonnées.

Keywords
big data | storage | big data | analytics

Abstract The purpose of this paper is to define the term of big data. First the big data are characterized through the model of 3V to 5V extended. The problematic of big data is distinguished from that of Business Intelligence. The economic and societal challenges associated with big data are discussed by presenting various examples of use in different areas of activity. Then, are introduced several great methods and techniques associated with storage and operation / analysis of these big data.

Pour toute question :
Service Relation clientèle
Techniques de l'Ingénieur
Immeuble Pleyad 1
39, boulevard Ornano
93288 Saint-Denis Cedex

Par mail :
infos.clients@teching.com
Par téléphone :
00 33 (0)1 53 35 20 20

Introduction au Big Data

Opportunités, stockage et analyse des mégadonnées

par **Bernard ESPINASSE**
Professeur des Universités,
Aix-Marseille Université,
École Polytechnique Universitaire de Marseille,
LSIS UMR CNRS 7296, Marseille, France.

et **Patrice BELLOT**
Professeur des Universités,
Aix-Marseille Université,
École Polytechnique Universitaire de Marseille,
LSIS UMR CNRS 7296, Marseille, France.

Depuis une vingtaine d'années, les données générées n'ont fait que s'accroître. Actuellement nous produisons annuellement une masse de données très importante estimée à près de 3 trillions (3.10^{18}) d'octets de données. On estime ainsi qu'en 2016 90 % des données dans le monde ont été créées au cours des deux années précédentes [3]. Selon le rapport IDC (International Data Corporation), la masse totale des données crée et copiée de par le monde pour 2011 était de 1,8 zettaoctets, soit de 10^{21} octets, et s'accroît d'un facteur 9 tous les 5 ans [15]. Cet accroissement des données touche tous les secteurs, tant scientifiques qu'économiques, ainsi que le développement des applications Web et les réseaux sociaux [14].

*Dans ce contexte, est apparu le terme **Big Data**. L'origine de ce terme anglo-saxon, littéralement « grosses données », est controversée, et sa traduction française officielle recommandée est **mégadonnées**, même si parfois on parle de **données massives**.*

Ces mégadonnées sont maintenant au centre des préoccupations des acteurs de tous les domaines d'activité. Ainsi le taux de croissance annuel moyen mondial du marché de la technologie et des services autour du Big Data sur la période 2011-2016 est estimé à plus de 30 %. D'après une étude IDC de 2013, ce marché devrait ainsi atteindre 23,8 milliards de dollars en 2016. Sur le plan européen, l'activité autour des mégadonnées devrait représenter autour de 8 % du PIB européen en 2020 (AFDEL février 2013). D'après le cabinet Markess International, le marché français des solutions et services en analytique, big data et gestion des données aurait atteint 1,9 milliard d'euros en 2015. Son taux de croissance annuel moyen d'ici 2018 est attendu à plus de 12 % (d'après Le monde informatique du 15 mars 2016).

*L'objet de cet article est de cerner ce terme Big Data ou mégadonnées, de préciser les enjeux économiques et sociétaux associés, d'introduire différentes grandes méthodes et techniques qui s'y rattachent. On s'intéresse dans cet article à deux grandes problématiques associées aux mégadonnées, d'une part leur stockage, les techniques traditionnelles de stockage de type bases de données relationnelles ne permettant pas de stocker de telles quantités de données, et d'autre part leur exploitation, l'analyse de ces mégadonnées dans des temps raisonnables. En effet, les mégadonnées s'accompagnent principalement du développement d'applications à visée analytique, qui traitent de données pour en tirer du sens. Ces analyses sont généralement appelées **Big Analytics**, ou **Analytique** ou encore **broyage de données**, reposant généralement sur des méthodes de calcul distribué.*

La section 1 présente une caractérisation du terme de Big Data ou Mégadonnées, en distinguant son paradigme de celui de l'informatique décisionnelle. Et quelques exemples d'usage des mégadonnées dans différents secteurs d'activité sont présentés à la section 2. La section 3 concerne la problématique du stockage de ces mégadonnées, tandis que la section 4 traite de la problématique de l'analyse des mégadonnées ou « analytique ».

1. Caractérisation des mégadonnées

Cette section présente une caractérisation des mégadonnées, notamment au travers du modèle populaire des « 3V » étendu au « 5V ». Ensuite une distinction est faite entre le paradigme de l'informatique décisionnelle et celui des mégadonnées.

1.1 Modèle des 3V étendu aux 5V

La caractérisation de ces mégadonnées ou Big Data est généralement faite selon 3 « V », les V de Volume, de Variété et de Vélocité, auxquels s'ajoutent d'autres « V » complémentaires, comme ceux de Valeur et de Véracité/Validité.

1.1.1 Volume

Le caractère « *volume* » est certainement celui qui est le mieux décrit par le terme « Big » de l'expression. Volume fait référence à la quantité d'informations, trop volumineuse pour être acquise, stockée, traitée, analysée et diffusée par des outils standards. Ce caractère peut s'interpréter comme le traitement d'objets informationnels de grande taille ou de grandes collections d'objets.

1.1.2 Variété

Le caractère « *variété* » fait référence à l'hétérogénéité des formats, de types, et de qualité des informations. Il est lié au fait que ces données peuvent présenter des formes complexes du fait qu'elles trouvent leurs origines dans des capteurs divers et variés (température, vitesse du vent, hygrométrie, tours/mn, luminosité...), dans des messages échangés (e-mails, médias sociaux, échanges d'images, de vidéos, musique), dans des textes, des publications en ligne (bibliothèques numériques, sites web, blogs...), des enregistrements de transactions d'achats, des plans numérisés, des annuaires, des informations issues des téléphones mobiles, etc.

1.1.3 Vélocité

Le caractère « *vélocité* » fait référence à l'aspect dynamique et/ou temporel des données, à leur délai d'actualisation et d'analyse. Les données ne sont plus traitées, analysées, en différé, mais en temps réel ou quasi réel. Elles sont produites en flots continus, sur lesquels des décisions en temps réel peuvent être prises. Ce sont les données notamment issues de capteurs, nécessitant un traitement rapide pour une réaction en temps réel. Dans le cas de telles données de grande vélocité engendrant des volumes très importants, il n'est plus possible de les stocker en l'état, mais seulement de les analyser en flux (*streaming*), voire de les résumer.

Ces 3V sont les caractéristiques majeures des mégadonnées. Aussi une définition plus pertinente des mégadonnées pourrait alors être : « des données qui sont trop volumineuses *ou* ayant une arrivée trop rapide *ou* une variété trop grande pour permettre de les ranger directement dans des bases de données traditionnelles ou de les traiter par les algorithmes actuels » [23].

1.1.4 Valeur

Le caractère complémentaire « *valeur* » fait référence à la potentialité des données, en particulier en termes économiques. Il est ainsi associé à l'usage qui peut être fait de ces mégadonnées, de leur analyse, notamment d'un point de vue économique. L'analyse de ces mégadonnées demande une certaine expertise tant liée à des méthodes et techniques en statistique, en analyse de données, que de domaine pour l'interprétation de ces analyses. Ainsi le McKinsey Global Institute avance que, dans les seuls États-Unis, il manquerait environ 150 000 personnes avec une expertise en analyse de *big data*. Cet organisme estime que le système de santé américain pourrait créer 300 milliards de dollars de valeur par an dont les deux tiers résulteraient en des réductions de coût d'environ 8 %. Les termes de « **Data Scientist** » et de « **Data Science** » sont liés à cette expertise recherchée et à cette nouvelle discipline émergente.

1.1.5 Véracité ou validité

Enfin, le caractère complémentaire « *véracité ou validité* » fait référence à la qualité des données et/ou aux problèmes éthiques liés à leur utilisation. Il comprend les problèmes de valeurs aberrantes ou manquantes (ces problèmes pouvant être résolus par le volume de données), mais aussi à la confiance que l'on peut avoir dans les données. S'il existe des critères permettant de qualifier la qualité des données, dans le cas de *big data*, cette vérification de la qualité est rendue difficile voire impossible du fait du volume, de la variété et de la vélocité spécifiques au Big Data.

1.2 Mégadonnées et informatique décisionnelle

Informatique décisionnelle et mégadonnées ont, toutes les deux, vocation à stocker et analyser des masses de données très importantes. La **Business Intelligence** (BI), traduite en français par « **informatique décisionnelle** » ou ID est apparue dans les années 1990 en management et en informatique. Elle a conduit à appréhender des données volumineuses principalement historisées et orientées sujet, stockées dans des entrepôts de données. À la fin des années 2000, a été introduit le terme de « **Business Analytics** » pour représenter la composante analytique clé dans l'ID [7]. Ces analyses étaient réalisées principalement par des **opérateurs d'analyse en ligne OLAP** (*On Line Analysis Processing*) sur des cubes extraits de ces entrepôts, ou par des techniques de **fouille de données** (*Data Mining*). Les données traitées en ID, stockées dans des entrepôts de données ou dans des cubes, sont des données multidimensionnelles : elles sont fortement structurées selon un modèle défini, de forte densité en information, et principalement numériques.

Ces analyses traitent des données afin de mesurer des phénomènes, notamment pour détecter des tendances. Elles s'appuient principalement sur la **statistique descriptive** ou **exploratoire**, dont l'objectif est de décrire des données à travers leur *présentation* (la plus synthétique possible), leur *représentation graphique*, et le calcul de *résumés numériques*. Dans cette optique, il n'est pas fait appel à des modèles probabilistes. Notons que si la statistique descriptive est intéressante, elle est coûteuse car elle repose sur des enquêtes portant sur un nombre important d'individus.

De leur côté, les mégadonnées concernent des données bien plus volumineuses que celles traitées par l'ID, structurées ou non, et de faible densité en information. Sur ces données de grand volume, elles utilisent la **statistique inférentielle** ou **inductive**, pour inférer des lois (régressions...). Les statistiques inférentielles utilisent la théorie des probabilités pour restreindre le nombre d'individus en faisant des sondages sur des échantillons. L'objectif principal est ainsi de préciser un phénomène sur une population globale, à partir de son observation sur une partie restreinte de cette population, l'échantillon. Il s'agit donc d'induire (ou encore d'inférer) du particulier au général avec un objectif principalement *explicatif*, ceci au moyen de modèles et d'hypothèses probabilistes. L'usage de la statistique inférentielle donne aux mégadonnées (avec les limites de l'inférence) des capacités prédictives [9]. De plus, la théorie des tests d'hypothèses permet de prendre des décisions dans des situations faisant intervenir une part de hasard.

Loin de les opposer, ID et mégadonnées peuvent s'enrichir l'une l'autre : l'ID doit apporter aux mégadonnées notamment ses méthodes de conception d'entrepôts et d'analyse, et les mégadonnées apportent leurs architectures de stockage distribuées et leurs analyses à larges échelles basées sur les statistiques inférentielles. Comme l'avance Delort [9], l'ID est plutôt basée sur un modèle du monde défini, alors que les mégadonnées visent à ce que les mathématiques (statistiques) trouvent un modèle dans les données.

2. De l'usage des mégadonnées

Les Mégadonnées ou Big Data sont dès à présent utilisées dans tous les secteurs d'activités, tant scientifiques, techniques que socio-économiques, depuis les données récupérées de l'exploitation de moteurs d'avion permettant de mieux maintenir ou concevoir ces derniers, jusqu'aux données spécifiant nos relations sur les réseaux sociaux pouvant être utilisées par les banques pour estimer la qualité de notre crédit... [9]. Donnons, de façon non exhaustive, quelques exemples d'usage des mégadonnées dans différents grands domaines d'activité.

2.1 Domaine de la recherche scientifique

Dans le domaine scientifique et technique, les scientifiques et ingénieurs font face à des mégadonnées notamment générées automatiquement par des capteurs ou instruments de mesure.

Par **exemple** dans le domaine de l'astronomie, en huit ans (2000-2008), le Sloan Digital Sky Survey, un grand programme d'observation astronomique, a enregistré 140 téraoctets d'images (140.10^{12}). Mais il ne faudra que cinq jours à son successeur, le LSST (Large Synoptic Survey Telescope) pour acquérir ce volume.

En physique, dans sa quête du boson de Higgs, le grand collisionneur de hadrons (LHC) a amassé de son côté, chaque année, près de 15 pétaoctets de données (15.10^{15}), l'équivalent de plus de 3 millions de DVD.

En recherche médicale, les technologies associées aux mégadonnées ont permis des avancées spectaculaires dans l'analyse du génome humain : alors qu'il a fallu dix ans, et plus de 2 milliards d'euros pour réaliser le premier séquençage humain complet, il est maintenant possible d'en réaliser un en quelques jours et pour environ mille euros. Ces connaissances sur le génome, couplées à d'autres, permettent de mieux comprendre l'évolution de pathologies, d'améliorer les mesures de prévention ou encore les protocoles de soins.

2.2 Domaine de la santé

Concernant le domaine de la santé, dans le rapport rendu public « *The big data revolution in healthcare* » [34], McKinsey évalue entre 300 et 450 milliards de dollars les sommes que le « big data » pourrait faire économiser au système de santé américain, sur un total de 2 600 milliards. Ces économies concernent notamment la prévention, avec un suivi des patients les incitant à changer leurs habitudes, le diagnostic, en aidant les médecins à choisir le traitement le plus approprié, le personnel médical, en déterminant si le patient a besoin d'une infirmière, d'un généraliste ou d'un spécialiste, la maîtrise des coûts, à la fois en automatisant les procédures de remboursement et en détectant les fraudes, et enfin l'innovation, à travers les multiples apports du calcul intensif à la compréhension du vivant et à l'amélioration des traitements.

De même grâce aux mégadonnées, il est possible de mieux prévenir certaines maladies ou épidémies, ou d'améliorer le traitement des patients. Ainsi en analysant les recherches des internautes sur Google, une équipe est parvenue à détecter plus rapidement l'arrivée des épidémies de grippe [12]. Autre exemple, en s'intéressant aux données disponibles sur Facebook, des chercheurs ont détecté les adolescents ayant des comportements à risque pour cibler les campagnes de prévention [22].

2.3 Domaine socio-économique et politique

Dans le domaine socio-économique, de façon générale, les mégadonnées peuvent être utilisées pour simplifier ou adapter des services offerts, ceci en écoutant mieux les usagers, en comprenant mieux leurs modes d'utilisation de ces services [13]. Ainsi Google Analytics propose par exemple aux entreprises, comme aux administrations publiques, d'améliorer la conception de leur site internet par l'analyse des visites des internautes.

Dans l'éducation, avec le télé-enseignement (dont les *Massive Open Online Courses* – MOOC), le traitement de mégadonnées permet d'analyser les activités des élèves (temps consacré, façon de suivre les programmes, arrêt-retour dans les vidéos pédagogiques, recherches Internet parallèles, etc.) pour améliorer les modes d'enseignement. L'analyse des mégadonnées permet aussi de mieux comprendre les sentiments ou les besoins des citoyens. Ainsi lors de la campagne de réélection de Barack Obama en 2012, les conseillers ont analysé en temps réel les messages sur Twitter pour adapter en direct le discours du président.

2.4 Domaine du transport et de l'énergie

Dans le domaine des transports, les déplacements des populations peuvent être modélisés pour adapter les infrastructures et les services (horaires des trains, etc.). À cette fin, les données provenant des *pass* de transports en commun, des vélos et des voitures partagées, mais aussi de la géolocalisation (données cellulaires et systèmes de localisation par satellites) de personnes ou de voitures, sont utilisées.

Dans le domaine de l'énergie et du développement durable, les systèmes de compteurs intelligents (électricité, gaz, eau) génèrent des mégadonnées qui permettent de rationaliser la consommation énergétique [23]. En plus d'offrir aux citoyens la possibilité de mieux contrôler leur consommation, ces compteurs permettent de couper à distance, avec l'accord des clients, l'alimentation d'équipements pour éviter les surcharges du réseau.

Dans le transport aérien, en associant les données issues de capteurs sur les avions à des données météo, on peut modifier les couloirs aériens pour réaliser des économies de carburant, on améliore la conception, la maintenance des avions ou leur sécurité [16].

3. Stockage et gestion des mégadonnées

Cette section traite de la problématique du stockage de très grands volumes de données. Dans un premier temps nous pointons les limites des bases de données relationnelles pour le stockage et la gestion des mégadonnées, et évoquons l'apport du *Cloud Computing* (informatique dans les nuages). Puis nous soulignons tout l'intérêt pour le stockage et la gestion des données du modèle de programmation parallèle « **MapReduce** » et du cadriciel libre « **Hadoop** » le mettant en œuvre. Ensuite nous introduisons les différents modèles de bases de données dites NoSQL, constituant différentes solutions de stockage des mégadonnées. Pour finir nous évoquons quelques autres alternatives, notamment les bases de données NewSQL.

3.1 Limites des bases de données relationnelles et *Cloud Computing*

En matière de stockage de données, les bases de données relationnelles restent la référence. Ces outils largement utilisés garantissent le maintien des propriétés ACID (Atomicité, Cohérence, Isolation et Durabilité). Pour gérer de gros volumes de données, notamment dans un contexte d'entrepôt de données, toujours fidèle au modèle relationnel, les machines bases de données, comme la Teradata™, s'appuient sur une distribution des données sur différents disques permettant une parallélisation de l'exécution des requêtes.

Cependant ces machines ne permettent de gérer des mégadonnées au-delà d'un certain volume. Aussi différentes nouvelles solutions ont vu le jour. Toutes ces solutions reposent sur un stockage distribué (partitionné) des données sur les clusters. Cependant, comme le théorème CAP de Brewer le démontre, aucun système distribué ne peut assurer à la fois la cohérence, la disponibilité et la possibilité d'être partitionné. La conséquence est que, dans ces nouvelles solutions de stockage, il ne sera pas possible d'assurer les propriétés ACID, et un relâchement de ces propriétés sera nécessaire.

Le nuage (*cloud*) est un ensemble de matériels, de raccordements réseau et de logiciels fournissant des services sophistiqués que des individus et des collectivités peuvent exploiter à volonté depuis n'importe où. Au lieu d'obtenir de la puissance de calcul par acquisition de matériel et de logiciel, dans le *Cloud Computing*, le consommateur utilise une puissance de calcul mise à sa disposition sur une architecture d'un fournisseur via Internet.

Le stockage des mégadonnées dans les nuages a tout son sens [1]. En effet, cette architecture est prévue pour passer à l'échelle horizontale notamment par une mutualisation de ressources hétérogènes. Les besoins de stockage s'accroissant, de nouveaux serveurs sont déployés dans cette architecture de façon transparente pour l'utilisateur. Si le *cloud* permet d'appréhender la caractéristique de *volume* des mégadonnées, les caractéristiques de *variété* et de *vélocité* ne le sont pas, le *cloud* étant davantage un support de stockage qu'une solution de gestion de données.

3.2 Intérêt de MapReduce et de Hadoop

Dans le stockage et la gestion des mégadonnées, le modèle de programmation parallèle « MapReduce » et le cadriciel libre « Hadoop » le mettant en œuvre s'avèrent déterminant.

MapReduce [8] est un patron ou modèle d'architecture de développement informatique, dans lequel sont effectués des calculs parallèles et souvent distribués sur des données pouvant être très volumineuses, par exemple supérieures en taille à 1 téraoctet. Il repose sur deux fonctions : « *Map* » et « *Reduce* », empruntées aux langages de programmation fonctionnelle. De façon générale, la fonction *Map*,

exécutée par un nœud spécifique, analyse un problème, le découpe en sous-problèmes, et ensuite délègue la résolution de ces sous-problèmes à d'autres nœuds de traitements pour être traités en parallèle, ceci à l'aide de la fonction *Reduce*. Ces nœuds font ensuite remonter leurs résultats au nœud qui les avait sollicités.

Ainsi le modèle MapReduce permet de manipuler de grandes quantités de données en les distribuant dans un cluster de machines pour être traitées. MapReduce a été rapidement utilisé par des sociétés intervenant sur le Web et possédant d'importants centres de traitement de données telles qu'Amazon ou Facebook. Notons que MapReduce est aussi de plus en plus utilisé dans le *Cloud Computing*. De nombreux cadriciels (*framework*) implémentant MapReduce ont vu le jour, dont le plus connu est Hadoop.

Hadoop (pour *High-Availability Distributed Object-Oriented Platform*), est un cadriciel (*framework*) de référence libre et *open source*, intégrant MapReduce et permettant d'analyser, stocker et manipuler de très grandes quantités de données. Hadoop a été créé par Doug Cutting et fait partie des projets de la fondation logicielle Apache depuis 2009.

Le noyau d'Hadoop est constitué d'une partie stockage consistant en un système de fichiers distribué, extensible et portable appelé HDFS (*Hadoop Distributed File System*), et d'une partie traitement appelée MapReduce. Hadoop fractionne les fichiers en gros blocs et les distribue à travers les nœuds du cluster. Pour traiter les données selon le modèle MapReduce, Hadoop transfère le code à chaque nœud et chaque nœud traite les données dont il dispose. Cela permet de traiter un volume important de données plus rapidement et plus efficacement que dans une architecture super-calculateur classique.

Les bases de données NoSQL, adaptées au stockage et à la gestion des mégadonnées, utilisent généralement le framework Hadoop pour analyser, stocker et manipuler ces mégadonnées. Notons enfin qu'ont vu le jour des *frameworks spécifiques* permettant d'améliorer les performances de Hadoop, notamment en milieu hétérogène, tant en termes de vitesse de traitement, qu'en termes de consommation électrique.

3.3 Bases de données NoSQL

Le stockage des mégadonnés nécessite un partitionnement. Or selon le théorème de Brewer (ou théorème de CAP pour *Consistency, Availability, Partition tolerance*), la gestion de ces mégadonnées partitionnées, est nécessairement supportée par un système distribué ne pouvant assurer simultanément la cohérence et la disponibilité des données, propriétés qu'assurent les systèmes relationnels (propriétés ACID). Aussi les systèmes de gestion de mégadonnées devront faire le choix entre cohérence et disponibilité. La disponibilité est alors généralement privilégiée dans l'exploitation de ces mégadonnées. De plus, les systèmes relationnels imposent une structuration des données selon des schémas spécifiques, or les mégadonnées sont en grande partie peu ou pas structurées.

En conséquence, de nouveaux modèles de stockage de données, mieux adaptés aux mégadonnées que le modèle relationnel, ont vu le jour, et ont conduit à l'émergence des **bases de données NoSQL**. Notons que le terme NoSQL, proposé par Carl Strozzi, ne signifie pas le rejet du modèle relationnel (SQL), mais doit être interprété comme « *Not Only SQL* ». Ces modèles ne remplacent pas les BD relationnelles mais sont une alternative, un complément apportant des solutions plus intéressantes dans certains contextes.

Ces systèmes NoSQL permettent une gestion d'objets complexes et hétérogènes sans avoir à déclarer au préalable l'ensemble des champs représentant un objet. Ils adoptent une représentation de données non relationnelle, sans schéma pour les données, ou avec des schémas dynamiques, et concernent des données de structures complexes ou imbriquées [28][21].

Les systèmes NoSQL apportent une plus grande performance dans le contexte des applications Web avec des volumétries de données exponentielles. Ils utilisent une très forte distribution de ces données et des traitements associés sur de nombreux serveurs (*sharding*). Ensuite ces systèmes optent pour un partitionnement horizontal des données sur plusieurs nœuds ou serveurs (*consistent hashing*). Enfin, ils utilisent généralement pour cela des algorithmes de type « MapReduce », pour paralléliser tout un ensemble de tâches à effectuer sur un ensemble de données.

En conséquence du théorème de Brewer, les systèmes NoSQL font un compromis sur le caractère « ACID » des systèmes relationnels : pour plus de scalabilité horizontale (passage à l'échelle) et d'évolutivité, ils privilégient la disponibilité à la cohérence, ils ne possèdent généralement pas de gestion de transactions. Pour assurer le contrôle de concurrence, ils utilisent des **mécanismes de contrôle multi-version** (MVCC) ou des **horloges vectorielles** (*Vector-Clock*). Ils sont principalement utilisés dans des contextes où il y a peu d'écritures et beaucoup de lectures.

L'adoption croissante des bases NoSQL par des grands acteurs du Web (Google, Facebook, Amazon, etc.), a conduit à une multiplication des offres de systèmes NoSQL, en grande partie en *Open-source*. Ces systèmes utilisent très souvent le *framework* Hadoop intégrant MapReduce et permettant d'analyser, stocker et manipuler de très grandes quantités de données. Enfin évoquons le système Hive construit sur Hadoop permettant d'écrire des requêtes dans le langage HQL (*Hive Query Language*) qui est proche du standard SQL.

3.4 Principaux modèles de bases de données NoSQL

Il existe pléthore de solutions NoSQL répondant plus ou moins bien à des besoins particuliers. Le but de cet article n'est pas de dresser un panorama complet des solutions existantes [5]. Cependant, ces approches peuvent globalement être regroupées en quatre catégories que nous décrivons succinctement dans ce qui suit [27].

3.4.1 Modèle orienté « clé-valeur »

Dans ce modèle (figure **1**), les données sont stockées sous la forme de grandes tables de hashage distribuées, permettant de facilement passer à l'échelle. Les données sont simplement représentées par un couple clé-valeur. La valeur peut être une simple chaîne de caractères, un objet sérialisé… Cette absence de structure et de typage a un impact important sur le requêtage. Ainsi la complexité d'une requête qui sera dans un système relationnel prise en charge par le langage SQL, sera dans ce modèle NoSQL prise en charge par l'applicatif interrogeant la base de données, la communication avec la base se limitant généralement à des ordres PUT, GET et DELETE.

Les systèmes NoSQL orientés clé-valeur les plus connus sont Memcached, Amazon's Dynamo, Redis, Riak et Voldemort créé par LinkedIn.

3.4.2 Modèle orienté « documents »

Ce modèle (figure **2**) se base sur le paradigme clé-valeur précédent ; cependant dans ce nouveau modèle la valeur est un document de type JSON ou XML. Ainsi, dans ce modèle, les données sont stockées à l'intérieur de documents. Un document peut être vu comme un n-uplet d'une table dans le monde relationnel, à la différence toutefois que les documents peuvent avoir une structure complètement différente les uns des autres. L'avantage est de pouvoir récupérer, via une seule clé, un ensemble d'informations structurées de manière hiérarchique. La même opération dans le monde relationnel impliquerait plusieurs jointures.

Les systèmes NoSQL orientés documents les plus connus sont CouchDB d'Apache, RavenDB (destiné aux plateformes .NET/Windows avec la possibilité d'interrogation via LINQ) et MongoDB.

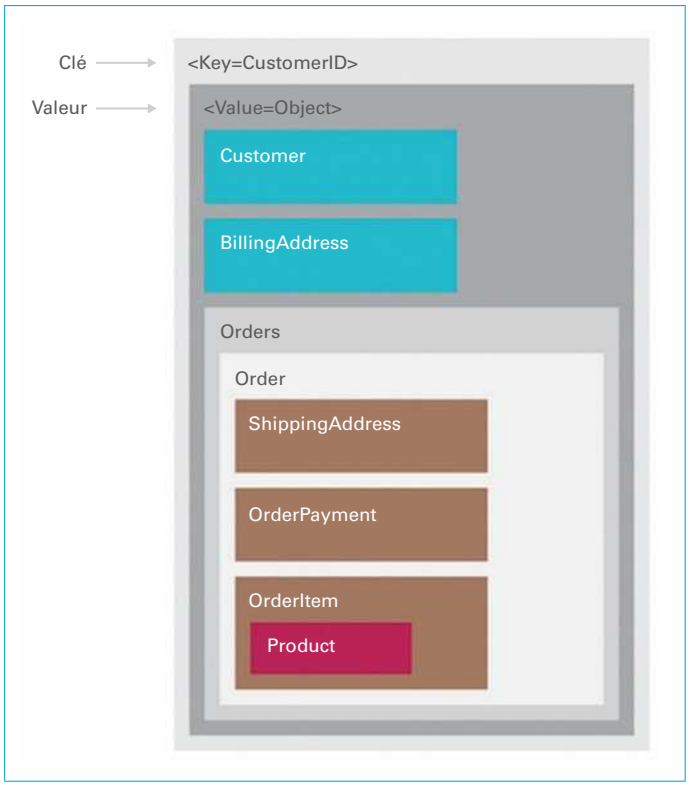

Figure 1 – Modèle NoSQL « clé-valeur » [27]

Figure 2 – Modèle NoSQL « document » [27]

3.4.3 Modèle orienté « colonnes »

Ce modèle (figure **3**) ressemble à première vue à une table du modèle relationnel, du fait que les attributs sont regroupés en famille de colonnes. Ainsi, deux attributs qui sont fréquemment interrogés ensemble seront stockés au sein d'une même famille de colonnes. Cependant la différence est que, dans cette base NoSQL orientée colonnes, le nombre de colonnes est dynamique, alors que dans une table relationnelle, le nombre de colonnes est fixé dès la création du schéma de la table. De plus, dans ce modèle, contrairement au modèle relationnel, le nombre de colonnes peut varier d'un enregistrement à un autre, ce qui évite de retrouver des colonnes ayant des valeurs inconnues (*Null Value*).

Les systèmes NoSQL orientés colonnes les plus connus sont principalement HBase, implémentation *Open Source* du modèle BigTable développé par Google, et Cassandra, projet Apache qui respecte l'architecture distribuée de Dynamo d'Amazon, et le modèle BigTable de Google.

3.4.4 Modèle orienté « graphe »

Ce modèle (figure **4**) qui repose sur la théorie des graphes, permet de représenter les données sous la forme de graphes. Le modèle s'appuie sur la notion de nœuds, de relations et de propriétés qui leur sont rattachées. Les entités sont alors les nœuds du graphe et les relations que partagent les entités sont alors des arcs qui relient

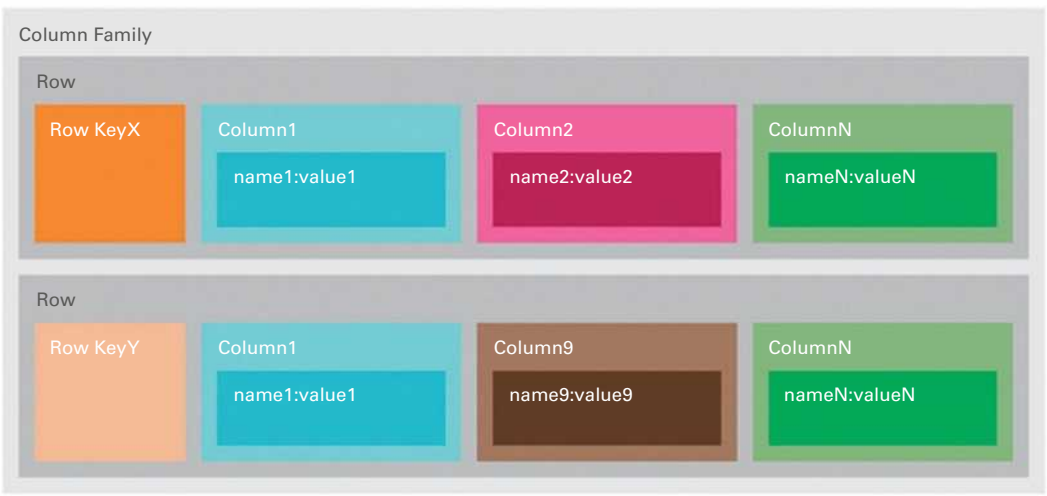

Figure 3 – Modèle NoSQL « colonnes » [27]

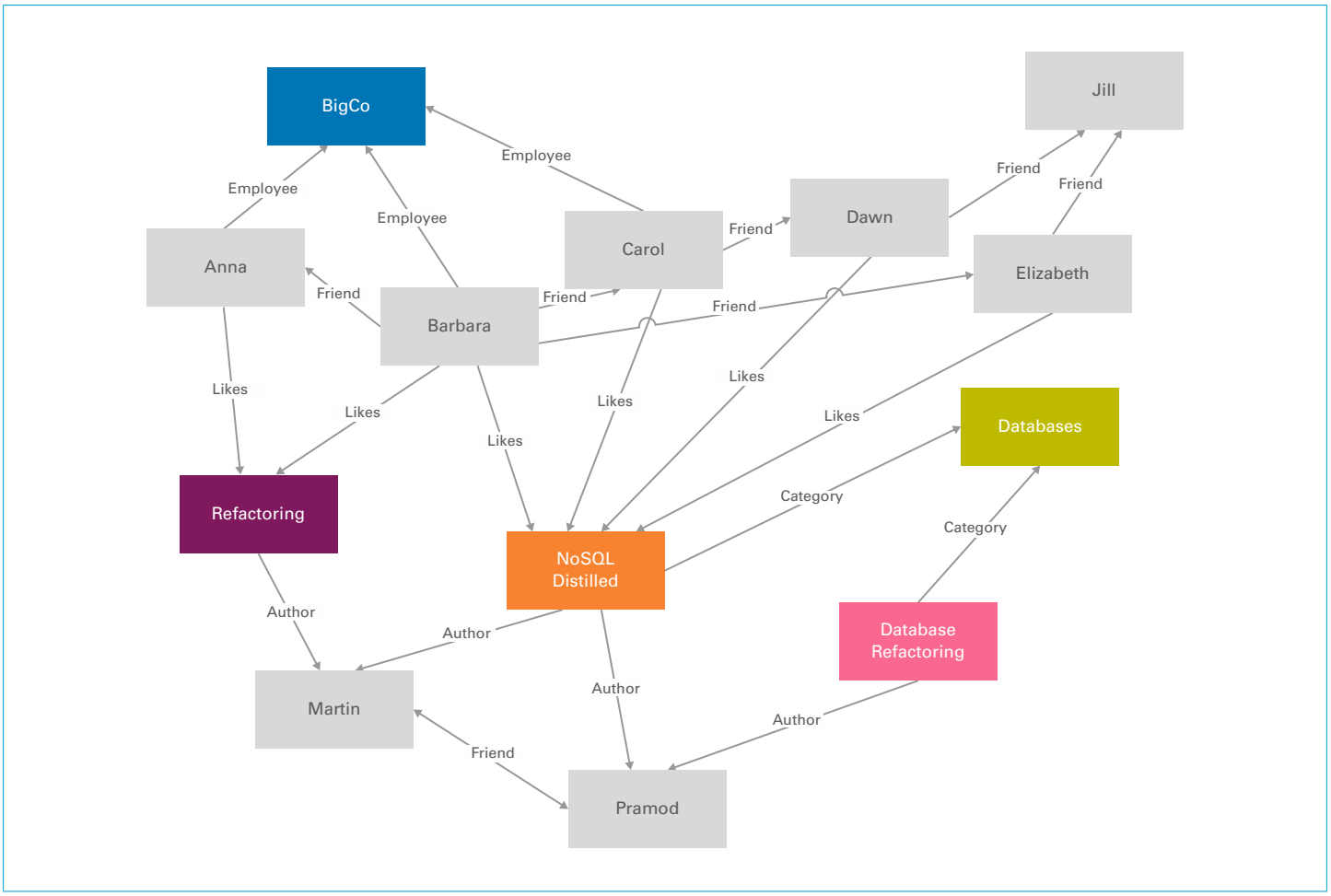

Figure 4 – Modèle NoSQL « graphe » [27]

ces entités. Ce modèle est notamment adapté au traitement des données des réseaux sociaux. Notons que les systèmes NoSQL orientés graphe trouvent un certain intérêt pour des applications dans le domaine du Web Sémantique, dans la gestion de bases de données de triplets RDF (*triple-stores*), permettant de stocker des connaissances ou ontologies, un triplet étant une arête d'un graphe.

Les systèmes NoSQL orientés graphe les plus connus sont Neo4J, Infinite Graph, OrientDB.

3.5 Alternatives au NoSQL : bases de données NewSQL

Pour pallier ces limitations et « réconcilier » le monde SQL et le monde NoSQL, de nouvelles architectures de stockage des données sont apparues récemment, regroupées sous le terme de **systèmes NewSQL**. Ces systèmes, encore en émergence, ont pour objectif une amélioration des performances des systèmes relationnels grâce à de nouveaux moteurs de stockage, des technologies transparentes de fragmentation, de nouveaux logiciels et matériels. Les systèmes NewSQL doivent permettre une interrogation des données via SQL tout en garantissant des performances et un passage à l'échelle similaires aux bases de données NoSQL. De plus ces systèmes se doivent de préserver les propriétés ACID. Parmi les systèmes NewSQL, on peut citer Clustrix, NuoDB, VoltDB ou encore F1, récemment proposé par Google.

Des produits Data Grid/Cache, sont aussi en émergence. Ils ont pour objectif une amélioration des performances notamment par un stockage des données persistantes en cache, une réplication des données distribuées, et un calcul exploitant le Grid.

Mais les systèmes relationnels n'ont pas dit leurs derniers mots. Ainsi Stonebraker et al. (2010) [26] comparent les systèmes NoSQL basés sur MapReduce aux systèmes relationnels parallèles commercialisés comme Terradata, Aster Data... Dans cette comparaison, ces derniers restent les plus performants au niveau du traitement des requêtes et des langages d'interrogation et interfaces qu'ils fournissent. Cependant avec MapReduce les systèmes NoSQL excellent dans les processus ETL (*Extract, Transform and Load*), et dans l'analyse d'ensembles de données semi-structurées en lecture seule.

Enfin l'émergence de moteurs analytiques basés sur MapReduce a un impact sur les produits systèmes relationnels parallèles commercialisés. Ainsi certains systèmes, comme Aster Data, permettent maintenant l'invocation de fonctions MapReduce sur les données stockées dans la base de données comme une partie d'une requête SQL. La fonction MapReduce apparaît dans la requête comme une table résultante à composer avec d'autres opérateurs SQL. D'autres éditeurs de systèmes relationnels fournissent des utilitaires pour déplacer des données entre des moteurs MapReduce et leurs moteurs relationnels, ce qui facilite notamment le passage des données structurées, distillées en analyse sur la plateforme MapReduce, dans le système relationnel.

4. Analyse des mégadonnées

Dans cette section, nous nous intéressons à la problématique de l'analyse de très grands volumes de données. La nature de cette analyse dépend de la nature et de la structure des mégadonnées, que l'on appelle aussi « analytique », traduction du terme anglo-saxon « *analytics* ». Ces différentes analyses mettront en œuvre divers algorithmes relevant de la fouille de données (*Data Mining*), de l'apprentissage machine automatique (*Machine Learning*), de l'aide à la décision, voire de la visualisation.

Nous soulignons tout d'abord l'intérêt de l'apprentissage automatique pour l'analyse de ces mégadonnées. Ensuite nous distinguerons l'analyse de mégadonnées stockées par exemple dans des systèmes NoSQL, et l'analyse de mégadonnées échangées et émises en continu, par exemple des données en flots, qu'il n'est pas envisageable de stocker du fait de leur volume. Les données concernées par les mégadonnées étant très diverses, de par leur nature et/ou leur niveau de structuration, leur analyse ou analytique sera différente. Aussi nous illustrerons quelques types d'analytiques associées à des grands types de mégadonnées : mégadonnées principalement composées de données numériques, mégadonnées textuelles, mégadonnées issues du Web, liées à des réseaux, et enfin liées aux mobiles. Comme nous l'évoquons plus loin, chacun de ces types d'analytique a ses caractéristiques propres et utilise des technologies plus ou moins spécifiques, et plus ou moins matures.

4.1 Intérêt de l'apprentissage automatique

Une part sans cesse croissante des recherches scientifiques et des développements logiciels est consacrée à l'apprentissage automatique. Cela s'explique par les succès de ces approches pour des tâches aussi diverses que la classification automatique de contenus, la fouille multimédia ou la compréhension du langage humain mais aussi par leurs grandes robustesses face à des données bruitées ou incomplètes.

D'une façon générale, l'apprentissage automatique consiste à déterminer automatiquement un modèle formel, décrivant les données disponibles et permettant un certain niveau de généralisation sur des données nouvelles.

Par **exemple**, pour une tâche de classification, il s'agira de déterminer un modèle (une fonction mathématique) mettant en correspondance l'espace de représentation des individus à classer (typiquement une liste de valeurs pour des propriétés choisies par un expert) et les classes cibles. Ce modèle est appris à partir d'exemples fournis au système après avoir été classés manuellement et selon un certain nombre d'hypothèses sur sa forme (par exemple un modèle linéaire si l'on suppose les données linéairement séparables).

L'objectif d'une méthode d'apprentissage est de déterminer le modèle qui minimise les erreurs de généralisation, c'est-à-dire qui permet d'obtenir une classification la plus exacte possible de données nouvelles. L'apprentissage s'effectuant sur un jeu de données réduit, toute la difficulté réside dans l'obtention d'un bon équilibre entre qualité de la classification sur les données exemples (ce sur quoi l'apprentissage est optimisé) et sur les données nouvelles, par définition inconnues au moment de l'apprentissage. Si l'apprentissage est essentiellement statistique via des modèles probabilistes dans les systèmes actuels, il peut aussi être symbolique par l'induction de règles, toujours à partir d'exemples, ou mixte comme c'est le cas avec les arbres de classification qui produisent, par analyse statistique, des règles de classification symboliques et compréhensibles par un opérateur humain. Par opposition, les méthodes d'apprentissage statistique produisent en effet des modèles « boîte noire », trop complexes pour être lisibles par un humain.

Relevons enfin que le coût de mise en œuvre des méthodes d'apprentissage tient d'une part dans la nécessité de disposer de données annotées manuellement du moins dans le cadre de l'apprentissage classique dit « supervisé ». Une des grandes nouveautés de ces dernières années est l'apprentissage par transfert qui exploite des données très diverses, non annotées, y compris pour apprendre des tâches spécifiques, en quantité plus ou moins grande selon les méthodes choisies, mais utilisé aussi le paramétrage des méthodes, leur test et la mise à jour des modèles. Hautement technique, la mise en œuvre de ces méthodes d'apprentissage suit un cheminement très différent des méthodes expertes traditionnelles qui souffrent d'être très spécifiques, peu robustes et surtout peu capables de passer à l'échelle des mégadonnées et de tirer profit de ces gisements nouveaux d'information.

4.2 Analyse de mégadonnées stockées

Comme nous l'avons vu au § 3, le stockage et l'exploitation des mégadonnées nécessitent une partition des données mais aussi une distribution des traitements, des algorithmes, nécessaires à l'accès et à la gestion de ces données. Il en sera de même pour les traitements et algorithmes d'analyse qui seront aussi distribués en utilisant ou non le *framework* MapReduce.

L'analyse des mégadonnées nécessite la mise en œuvre de traitements et d'algorithmes, notamment d'apprentissage automatique et de fouille de données, qui doivent aussi être distribués pour être efficaces. Ainsi sur la base d'Hadoop, s'est développé le projet *Mahout* fournissant des versions distribuées de plusieurs algorithmes standards d'apprentissage automatique et de fouille de données, comme des algorithmes de factorisation de matrices, utilisés par exemple dans les systèmes de recommandation, des algorithmes de classification tels que les k-moyennes qui permettent d'organiser une collection de documents (ou plus généralement d'objets représentés sous formes de vecteurs) en classes ou encore des algorithmes de catégorisation tels que les forêts aléatoires (*Random Forest*) ou les classifieurs bayésiens.

La mise en œuvre de ces algorithmes distribués d'analyse de données, incluant la fouille, l'apprentissage, l'aide à la décision, et la visualisation, nécessite de les réécrire pour en proposer une version distribuée, mais aussi d'utiliser des environnements matériels spécifiques permettant de les exécuter en mode distribué, par exemple des machines multi-cœurs ou des grilles de calcul. Les limites actuelles à la distribution des algorithmes d'analyse des mégadonnées, l'analyse prise ici au sens large, incluant la fouille, l'apprentissage, l'aide à la décision, la visualisation, résident tant dans la difficulté algorithmique, que dans l'architecture matérielle d'exécution.

Ainsi la plupart des algorithmes de fouille ne se distribuent pas facilement, et pas nécessairement avec une approche de type MapReduce, et il est parfois nécessaire de trouver de nouvelles techniques pour réaliser une parallélisation efficace [23]. Ces auteurs soulignent aussi le manque de langages standardisés devant permettre aux développeurs d'utiliser facilement les implantations parallèles existantes et d'en proposer de nouvelles.

4.3 Analyse de flots de données

Les mégadonnées ne concernent pas seulement les données stockées, par exemple dans des systèmes NoSQL. Elles concernent aussi les données échangées et émises en continu, comme des données en flots sur des médias en ligne, des données en provenance de capteurs, ou encore des relevés d'expérimentation dans le domaine de la physique.

Lorsqu'il s'agit de requêter un flux de données continu, rapide et sans fin, il n'est pas envisageable d'interroger la totalité du flux,

ce qui pourrait avoir pour conséquence de stopper le flux. De nouveaux algorithmes ont donc été optimisés en temps de traitement, et en occupation mémoire, pour répondre à cette contrainte d'exploration et d'analyse de données.

Parmi les techniques les plus utilisées dans la fouille de flots de données, citons celles permettant de construire des résumés ou synopsis. Ces techniques n'explorent pas le flot entier, mais interrogent des données sélectionnées dans le flux, on accepte ainsi des résultats avec une certaine approximation. Les fenêtres temporelles sont une de ces techniques travaillant sur un ensemble restreint du flux pour en extraire des motifs (items, itemsets, et motifs séquentiels ou *Sequential patterns*) porteurs de connaissance [23]. D'autres techniques comme les histogrammes, la compression, les sketches, l'échantillonnage statistique, permettent aussi de créer des résumés de flux de données et d'effectuer des fouilles sur ces résumés. Les résumés servent aussi à « **ralentir** » le flot de données quand il est trop rapide pour les machines effectuant l'analyse.

Les grands défis que posent les flots de données résident tout d'abord dans le fait que les algorithmes nécessaires pour les traiter ne disposent que d'un espace mémoire réduit, et qu'il n'est possible de stocker qu'une faible proportion des données reçues. Ensuite ces algorithmes disposent d'un temps limité pour effectuer les traitements voulus, ils ne peuvent généralement effectuer qu'une seule passe sur les données. Cette complexité augmente si l'on cherche des résumés des données plus riches ou si les données possèdent une structure complexe, par exemple sous forme de graphes, comme dans les réseaux sociaux. Ainsi le déploiement d'algorithmes de prédiction de diffusion, dans de tels graphes pose d'importants problèmes algorithmiques [23].

La nature dynamique des données en ligne a de plus donné un nouvel essor aux travaux sur les séries temporelles, et l'on assiste depuis quelques années à la proposition de nouvelles méthodes pour réaliser des tâches de classification, supervisée ou non, et de prédiction sur les séries temporelles. Elle a aussi remis en avant les travaux sur la construction incrémentale de modèles, ainsi que ceux liés à la visualisation des données [H 7 418], le défi étant ici de proposer des méthodes de visualisation qui permettent d'avoir un bon aperçu des données traitées sans trop sacrifier à la qualité du résumé proposé.

4.4 Analyse de données

L'analyse des données (*Data Analytics*) repose principalement sur la fouille de données (*Data Mining*) et sur l'analyse statistique. La plupart de ces techniques reposent sur les technologies commerciales matures que sont les SGBD relationnels, les entrepôts de données, les ETL, l'analyse OLAP et l'analyse des processus.

Depuis la fin des années 1980, les chercheurs en IA (Intelligence Artificielle), en algorithmique, et en base de données ont développé divers algorithmes d'exploration de données. Dans la conférence internationale ICDM 2006 en fouille de données, les dix algorithmes d'exploration de données les plus importants ont été identifiés, ceci sur la base d'experts, de nombre de citations, et sur une enquête communautaire. Par ordre d'importance décroissante on trouve les arbres de décision (C4.5), la classification par k-moyenne (clustering de type k-means), la classification supervisée par approche discriminante et fonctions noyaux de type SVM (*Support Vector Machine*), Apriori, l'estimation automatique de distribution de probabilités par approche EM (*Expected Maximisation*), l'estimation de scores d'autorité ou de popularité par marche aléatoire au sein de graphes comme l'algorithme Google PageRank, la classification au moyen de classifieurs simples mais « boostés » comme AdaBoost, la catégorisation par calcul de similarités avec la méthode des kNN (k plus proches voisins), la classification probabiliste bayésienne naïve (*Naive Bayes*), et les arbres de classification CART [32].

Ces algorithmes couvrent

– la **classification** qui est l'attribution automatique d'un individu à une classe pré-existante ;

– le **clustering**, regroupement automatique d'individus au sein d'un certain nombre de classes a priori inconnues ;

– la **régression** qui est une estimation automatique d'une fonction mathématique permettant de faire correspondre des entrées, par exemple des vecteurs décrivant des individus et des sorties, ou encore des classes ou des valeurs numériques ;

– l'**analyse d'association**, entre individus ou entre variables ;

– l'**analyse de réseaux**, ou graphes.

Notons que la plupart de ces algorithmes de fouille de données sont maintenant intégrés dans des systèmes de fouille de données commercialisés (Matlab notamment) ou *open source* (tels que Weka [31], l'environnement R ou SciKit pour Python). À ces algorithmes de base, il faut rajouter les développements récents consécutifs de la remise au goût du jour des approches d'apprentissage automatique par réseaux de neurones. Couplés en sortie à des classifieurs de type régression linéaire ou logistique ou à des classifieurs bayésiens, ils permettent d'apprendre des représentations riches d'individus en très grande quantité.

Issu de l'approche connexionniste (réseaux de neurones – réseaux convolutifs), l'**apprentissage profond** (*Deep Learning*) est actuellement l'objet d'un fort engouement des communautés scientifiques et d'ingénieurs qui proposent de nombreuses variantes et des architectures ouvertes (par exemple TensorFlow de Google). Difficile à paramétrer (architectures récurrentes ou non, nombre de couches, taille des vecteurs représentant les individus, fonctions d'activation...) et nécessitant un très grand nombre de données en exemple, l'apprentissage profond a cependant produit des résultats significativement meilleurs que toutes les autres méthodes pour la fouille d'images, la reconnaissance de la parole ou de caractères. Sur le plan technique, ces méthodes peuvent être mises en œuvre de façon massivement parallèle via l'exploitation de cartes GPU.

Du fait des succès obtenus par les communautés de chercheurs en fouille de données et en analyse statistique, l'analyse des données continue d'être un domaine de recherche très actif, principalement basé sur des modèles mathématiques bien fondés et des algorithmes puissants, des techniques telles que les réseaux bayésiens, les modèles de Markov cachés (HMM), les machine à vecteur support (SVM), et les modèles d'ensemble. L'apprentissage statistique a été appliqué à de l'analyse de données, de textes et du Web [6].

Au-delà, c'est l'approche générale qui fait aussi l'objet de variantes plus ou moins efficaces selon les contextes et les données disponibles et manipulées. Supervisé (à base d'exemples pré-étiquetés ou classés), semi-supervisé (combinaison d'exemples étiquetés et d'exemples non étiquetés pour l'apprentissage) ou non supervisé (apprentissage d'un modèle à partir des seules données non étiquetées), l'apprentissage peut aussi être actif (sélection automatique des meilleurs exemples à étiqueter manuellement parmi un ensemble très important de données non étiquetées) ou par renforcement (le système apprend au fur et à mesure que l'on s'en sert, à la manière d'un joueur qui gagne en expérience).

D'autres techniques d'analyse de données ont vu le jour pour explorer des données spécifiques, de la fouille de données séquentielles, temporelles ou spatiales, à la fouille de flux de données à haut débit et les données de capteurs. Les préoccupations croissantes de respect de la vie privée dans diverses applications d'e-commerce, e-gouvernement, et de santé ont conduit à l'émergence de techniques de fouille de données respectant cette vie privée (*privacy-preserving data mining*), utilisant généralement des techniques d'anonymisation, qui sont des méthodes dirigées par les données, ou des méthodes dirigées par les processus défi-

nissant comment les données peuvent être accédées et utilisées [11].

Citons aussi la fouille de processus (*process mining*), basée sur l'analyse de données événementielles, et permettant la découverte de nouveaux processus ou le contrôle de conformité de processus, en exploitant des journaux d'événements (*event logs*) de plus en plus disponibles dans les organisations quel que soit le domaine, de l'industrie à la santé [30].

4.5 Analyse de textes

Une partie importante du contenu non-structuré recueilli par une organisation est en format textuel, qu'il s'agisse de la communication par e-mail et des documents d'entreprise ou de pages Web et du contenu des médias sociaux. L'**analyse de texte** (*Text analytics*) relève de la **recherche d'information** (RI), de la **fouille de texte** (*Text Mining*) et de la **linguistique informatique**. Dans la RI, la représentation des documents et le traitement des requêtes sont les fondements de l'élaboration du modèle vectoriel, du modèle booléen, et du modèle probabiliste, qui sont à la base de l'exploitation des bibliothèques numériques modernes, des moteurs de recherche et des systèmes de recherche d'entreprise [25].

En linguistique informatique, on dispose maintenant de techniques matures de **traitement automatique du langage naturel** (TALN ou NLP – *natural language processing*), principalement statistiques, mais aussi symboliques, pour l'acquisition lexicale et l'extraction de terminologie, la désambiguïsation sémantique, la reconnaissance d'entités nommées (noms propres, dates, quantités...), l'étiquetage syntaxique [18]. En plus de la représentation de documents et de requêtes, des modèles d'utilisateur par retour de pertinence tenant compte des comportements et de l'expression de jugements (*relevance feedback*) sont également utilisés dans l'amélioration des performances de recherche. Notons que les moteurs de recherche actuels utilisent de plus en plus des techniques de TALN pour réduire l'impact de l'utilisation d'un vocabulaire varié entre les requêtes et les documents, pour exploiter des requêtes plus complexes ou tout naturellement pour mieux appréhender la sémantique des contenus.

Tirant parti de la puissance des mégadonnées en apprentissage, et du TALN statistique pour construire des modèles de langue (distributions de probabilités décrivant comment les mots apparaissent les uns par rapport aux autres), les techniques d'analyse de textes ont été utilisées dans plusieurs domaines émergents, comme l'extraction d'information, les systèmes de question-réponse (question en langue naturelle et extraction automatique des réponses au sein de de documents textuels) et en analyse d'opinions et de sentiments [6].

L'**extraction d'information** vise à extraire automatiquement des types spécifiques d'informations à partir de documents. Elle peut être vue comme un moyen de structurer automatiquement des phrases ou des documents. La tâche de **reconnaissance d'entités nommées** (REN ou NER pour *named entity recognition*) est un processus qui identifie les éléments atomiques dans le texte, et les classe en catégories prédéfinies (par exemple, noms, lieux, dates). Les techniques de NER ont été développées avec succès notamment pour l'analyse des nouvelles (*news*) et dans l'extraction d'information dans le domaine biomédical. Actuellement les méthodes statistiques permettent d'extraire plus de 90 % des entités nommées. Cependant l'extraction de relations entre ces entités nommées est plus difficile, nécessite plus de sémantique, surtout lorsqu'il s'agit de relations n-aires (extraction d'événements).

Les « *topic models* » correspondent à une famille d'algorithmes permettant de découvrir les principaux thèmes qui imprègnent une grande collection non structurée de documents. Dans ce cas, un thème n'est pas représenté par un mot mais par un vecteur de mots, chacun associé à un score.

Par **exemple**, l'allocation de Dirichlet latente ***(LDA – Latent Dirichlet Allocation)*** [2] s'appuie sur une approche générative (modélisation de probabilités jointes et non pas conditionnelles). Elle pose l'hypothèse que chaque thème est associé à une distribution probabiliste de mots (une fois un thème choisi, il est possible de ***générer*** un texte à son sujet) mais aussi que chaque document correspond à une distribution probabiliste de thèmes (il est possible de générer un document comme étant une mixture de thèmes). Utilisée en analyse de documents, la LDA permet d'obtenir la liste des thèmes (une liste de listes de mots) qui les caractérisent au mieux.

Plusieurs variantes ont été proposées, parmi lesquelles les *Author Topic Models*, qui permettent d'identifier les thèmes qui caractérisent les auteurs et les *Sentiment Topic Models* qui proposent de différencier les thèmes qui font l'objet de sentiments positifs des autres.

Les **systèmes Question-Réponse** (*Question answering systems*) reposent sur des techniques du TALN, de la RI, et l'interaction homme-machine (IHM). Principalement conçus pour répondre à des questions factuelles (de type « qui, quoi, quand et où »), ces systèmes impliquent des techniques différentes pour l'analyse de la question exprimée par l'utilisateur en langue naturelle, la recherche de la source (quels sont les documents les plus susceptibles de contenir les réponses cherchées), l'extraction de la (ou des) réponse(s) et leur présentation aux utilisateurs [20]. Les récents succès de Watson d'IBM et de Siri d'Apple ont mis en évidence le potentiel de ces systèmes Question-Réponse ainsi que des opportunités de commercialisation dans de nombreux domaines d'application, notamment l'éducation, la santé, et la défense [6].

La **fouille ou analyse d'opinion** ou de sentiments (*sentiment analysis*) se réfère aux techniques pour extraire, classer, comprendre et évaluer les sentiments exprimés dans diverses sources en ligne, dans des commentaires sur les médias sociaux, et dans d'autres contenus générés par les utilisateurs [H 7 270]. L'analyse de sentiments est l'objet de plusieurs variantes destinées à estimer l'affect, la subjectivité, et d'autres états émotionnels dans les textes en ligne. Le Web 2.0 et le contenu des médias sociaux ont créé de nombreuses opportunités passionnantes pour comprendre les opinions du grand public et des consommateurs en ce qui concerne les événements sociaux, les mouvements politiques, les stratégies d'entreprise, les campagnes de marketing, et les préférences de produits [24].

L'**analyse de textes** offre également des opportunités et des défis de recherche importants dans plusieurs domaines plus ciblés, y compris l'analyse web « stylométrique » pour l'attribution d'auteur (détection de plagiat), l'analyse multilingue pour les documents web, et la visualisation à grande échelle de collections [6], par exemple l'environnement logiciel TXM (textométrie) ou Gephi pour les graphes. N'oublions pas le résumé automatique de textes, avec notamment ses approches extractives utilisant des techniques statistiques, appliquées à un seul document ou à un ensemble de documents [17][29][H 7 260].

Dans le domaine multimédia, ces approches de fouille de texte sont exploitées via des modules de transcription automatique de la parole en texte, de reconnaissance du locuteur pour distinguer les intervenants dans un flux audio et structurer le document en sortie ou d'analyse d'images pour différencier des plans dans une vidéo. La prise en compte de critères extra-linguistiques (gestuelle, prosodie...) dans l'analyse de contenus multimédia fait l'objet de recherches intensives partant du constat que la compréhension humaine du langage ne peut être pleinement dissociée de son environnement.

Enfin, comme pour l'analyse de données, l'analyse de textes tire profit d'implémentations orientées « Big Data » autour de MapReduce et d'Hadoop, des services du *cloud*, des bases NoSQL (par exemple Apache Solr) et des modules matériel de type GPU (par exemple Word2Vec).

4.6 Analyse du Web

Depuis près de dix ans, l'analyse du Web (*Web Analytics*) constitue un thème de recherche très actif avec de nombreux challenges et opportunités. L'analyse du Web regroupe les méthodes et technologies relatives à la collecte, la mesure, l'analyse et la présentation des données utilisées dans les sites et applications Web [33][4].

L'analyse du Web n'a cessé de croître, et est passée d'une simple fonction HTTP de journalisation du trafic, à une suite plus complète d'outils permettant le suivi, l'analyse et la création de rapports sur des données d'utilisation sur le Web. L'industrie et le marché de l'analyse du Web sont en plein essor.

Elle s'appuie principalement sur les avancées en fouille de données, en recherche d'information (RI), et en traitement automatique des langues naturelles (TALN) déjà utilisé en l'analyse de textes comme vu précédemment.

Les sites Web basés sur http et html inter-reliés, les moteurs de recherche sur le Web, et les systèmes d'annuaire pour localiser le contenu Web, ont contribué à développer des technologies spécifiques pour l'exploration des sites sur le Web (robots d'indexation – *web crawler* ou *web spider*), la mise à jour des pages Web, le classement des sites Web, et l'analyse des logs de recherche, maintenant intégrés dans les systèmes de recommandation [H 7 245]. Cependant, l'analyse du Web est devenue encore plus excitante avec la maturité et la popularité des services Web et l'arrivée des systèmes du Web 2.0 dans le milieu des années 2000.

Basés sur XML et les protocoles Internet http et smtp, les services Web offrent une nouvelle façon de réutiliser et d'intégrer des systèmes tiers ou existants. De nouveaux types de services Web et leurs API associées (interface de programmation d'application) permettent aux développeurs d'intégrer facilement des contenus divers issus de différents « *web-enabled* » systèmes.

Citons par **exemple** REST (*Representational State Transfer*) pour invoquer des services à distance, RSS (*Really Simple Syndication*) pour le « *pushing* » de nouvelles, JSON (*JavaScript Object Notation*) pour les échanges légers de données, et enfin AJAX (*asynchronous JavaScript + XML*) pour l'échange de données et d'affichage dynamique.

Ces modèles de programmation légers permettent la syndication et la notification de données, ainsi que les agrégations (*mashups*) de contenus multimédia (Flickr, Youtube, Google Maps) à partir de différentes sources, un processus web similaire au processus ETL (*Extraction, Transformation and Loading*) des entrepôts de données. La plupart des fournisseurs d'e-commerce ont fourni des API matures pour accéder à leur produit et au contenu de leur clientèle.

Par **exemple**, grâce à Amazon Web Services, les développeurs peuvent accéder au catalogue de produits, commentaires des internautes, classement du site, les historiques de prix, et au Amazon Elastic Compute Cloud (EC2) pour calculer la capacité. De même, les API Web de Google prennent en charge la recherche AJAX, Map API, GData API (pour Calendar, Gmail...), Google Translate, et Google App Engine pour les ressources de ***cloud computing***.

Les services Web et les API continuent de fournir un flux impressionnant de nouvelles sources de données pour les mégadonnées.

Une importante composante émergente dans la recherche en analyse du Web est le développement de plateformes et services de *cloud computing*, qui comprennent des applications, des logiciels système et du matériel fournis comme services sur Internet. Basé sur une architecture orientée services (SOA), sur la virtualisation des serveurs et sur l'informatique utilitaire (*utility computing*), le *cloud computing* peut être offert en tant que logiciel comme service (SaaS), l'infrastructure en tant que service (IaaS), ou plateforme en tant que service (PaaS) [H 6 020].

Par **exemple**, Amazon Elastic Compute Cloud (EC2) permet aux utilisateurs de louer des ordinateurs virtuels, sur lesquels ils peuvent exécuter leurs propres applications informatiques. Son Simple Storage Service (S3) offre un service de stockage en ligne de Web. Google App Engine fournit une plateforme pour le développement et l'hébergement Java ou des applications Web basées sur Python. Google Bigtable est utilisé pour le stockage de données d'arrière-plan (*backend*). La plateforme Windows Azure de Microsoft fournit des services de *cloud computing* tels que SQL Azure et SharePoint, et permet aux applications en. NET de fonctionner sur la plateforme.

Les recherches en analyse du Web englobent maintenant la recherche et la fouille sociale, les systèmes de réputation, l'analyse des médias sociaux, et la visualisation Web. De plus, les ventes aux enchères sur le Web, la monétisation d'Internet, le marketing social et la confidentialité/sécurité du Web sont quelques-uns des axes de recherche prometteurs liés à l'analyse du Web. Beaucoup de ces nouveaux domaines de recherche peuvent compter sur les progrès dans l'analyse des réseaux sociaux, l'analyse de texte, et même dans la recherche en modélisation économique.

5. Conclusion

Cet article a permis de cerner ce terme de « Big Data » ou mégadonnées, actuellement au centre des préoccupations des acteurs de tous les domaines d'activité, en évoquant les principaux enjeux économiques et sociétaux associés. Bien des aspects n'ont pas été traités dans cet article, citons notamment les problèmes éthiques liés à leur utilisation.

Nous avons aussi introduit les différentes grandes méthodes et techniques qui s'y rattachent, tant en ce qui concerne leur stockage, que leur exploitation de par leur analyse. Ces méthodes et techniques, tout comme les outils logiciels qui y sont rattachés sont encore en devenir.

Avec un taux de croissance annuel moyen mondial de près de 30 % du marché de la technologie et des services autour des mégadonnées, un besoin très important en expertises humaines se fait sentir, expertises plus particulièrement liées à l'analyse de ces mégadonnées, et relevant d'une discipline en pleine émergence qu'est la science des données (*Data Science*).

Introduction au Big Data

Opportunités, stockage et analyse des mégadonnées

par **Bernard ESPINASSE**

Professeur des Universités,
Aix-Marseille Université,
École Polytechnique Universitaire de Marseille,
LSIS UMR CNRS 7296, Marseille, France.

et **Patrice BELLOT**

Professeur des Universités,
Aix-Marseille Université,
École Polytechnique Universitaire de Marseille,
LSIS UMR CNRS 7296, Marseille, France.

Sources bibliographiques

[1] AGRAWAL (D.), DAS (S.) et EL ABBADI (A.). – *Big data and cloud computing : current state and future opportunities*. In Proceedings of the 14th International Conference on Extending Database Technology (pp. 530-533). ACM (2011).

[2] BLEI (D.). – Probabilistic Topic Models, Communications of the ACM (55 : 4), pp. 77-84. M (2012).

[3] BRASSEUR (C.). – *Enjeux et usages du big data*. Technologies, méthodes et mises en œuvre, Paris, Lavoisier, p. 30 (2013).

[4] BURBY (J.) et BROWN (A.). – *Web Analytics Definitions – Version 4.0*. Retrieved from http://www.digitalanalyticsassociation.org/ standards (2007).

[5] CATTELL (R.). – *Scalable SQL and NoSQL data stores*. ACM SIGMOD Record, 39 (4), pp. 12-27 (2011).

[6] CHEN (H.), CHIANG (R.H.L.) et STOREY (V.C.). – *Business Intelligence and Analytics : From Big data to Big Impact*, MIS Quarterly, Vol. 36 No. 4, pp. 1165-1188/December 2012 (2012).

[7] DAVENPORT (T.H.). – *Competing on Analytics*, Harvard Business Review (84 : 1), pp. 98-107 (2006).

[8] DEAN (J.) et GHEMAWAT (S.). – *MapReduce : Simplified Data Processing on Large Clusters*, OSDI 2004 (2004).

[9] DELORT (P.). – *Le Big Data*. Presses Universitaires de France (2015).

[10] GAMBHIR (M.) et GUPTA (V.). – *Recent automatic text summarization techniques : a survey*, Artificial Intelligence Review, March 2016, DOI 10.1007/s10462-016-9475-9 (2016).

[11] GELFAND (A.). – *Privacy and Biomedical Research : Building a Trust Infrastructure*. An Exploration of Data-Driven and Process-Driven Approaches to Data Privacy, Biomedical Computation Review, Winter, pp. 23-28 (2011).

[12] GINSBERG (J.) et al. – *Detecting influenza epidemics using search engine query data*. Nature, n° 457, pp. 1012-1014 (2009).

[13] HAMEL (M.-P.) et MARGUERIT (D.). – *Analyse des big data usages, quels défis ?* Note d'analyse du Commissariat général à la stratégie et à la prospective, N° 8, nov. 2013 (2013).

[14] HELBING (D.) et POURNARAS (E.). – *Build Digital Democracy : Open Sharing of Data that are Collected with Smart Devices would Empower Citizens and Create Jobs*.Nature, Vol. 527, Nov. 2015, Macmillan Publishers (2015).

[15] IDC-2011, GANTZ (J.) et REINSEL (D.). – *Extracting Value from Chaos*. IDC iView, pp. 1-12 (2011).

[16] JOUNIAUX (P.). – *Big data au service de la sécurité du transport aérien : l'analyse des données de vol*, Télécom, n° 169, juillet (2013).

[17] LLORET (E.) et PALOMAR (M.). – *Text summarisation in progress : a literature review*. Artif. Intell. Rev., 37 (1), pp 1-41 (2012).

[18] MANNING (C.D.) et SCHÜTZE (H.). – *Foundations of Statistical Natural Language Processing*, Cambridge, MA : The MIT Press (1999).

[19] MAREK (K.). – *Web Analytics Overview*. Using Web Analytics in the Library, Lybrairy Technology Reports, alatechsource.org, July 2011 (2011).

[20] MAYBURY (M.T.) (ed.). – *New Directions in Question Answering*, Cambridge, MA : The MIT Press. (2004).

[21] MONIRUZZAMAN (A.B.M.) et HOSSAIN (S.A.). – *NoSQL Database : New Era of Databases for Big data Analytics – Classification, Characteristics and Comparison*, International Journal of Database Theory and Application, Vol. 6, No. 4, 2013 (2013).

[22] MORENO (M.) et al. – *Associations between displayed alcohol references on Facebook and problem drinking among college students*, Archives of Pediatrics & Adolescent Medicine, 166 (2), pp. 157-163 (2012).

[23] MOTHE (J.), PITARCH (Y.) et GAUSSIER (E.). – *Big Data : Le cas des systèmes d'information*, Revue Ingénierie des Systèmes d'Information, Hermès Éditeur, Vol. 19/3 (2014).

[24] PANG (B.) et LEE (L.). – *Opinion Mining and Sentiment Analysis*, Foundations and Trends in Information Retrieval (2:1-2), pp. 1-135 (2008).

[25] SALTON (G.). – *Automatic Text Processing*, Reading, MA, Addison Wesley (1989).

[26] STONEBRAKER (M.), ABADI (D.), DEWITT (D.J.), MADDEN (S.), PAVLO (A.) et RASIN (A.). – *MapReduce and Parallel DBMSs : Friends or Foes*, Communications of the ACM (53 : 1), pp. 64-71 (2012).

[27] SADALAGE (P.). – *NoSQL Databases : An Overview*. Source : https:// www.thoughtworks.com/insights/blog/ nosql-databases-overview

[28] STRAUCH (C.). – *Nosql databases*, Lecture Notes, Stuttgart Media University (2011).

[29] TORRES-MORENO (J.-M.). – *Résumé automatique de documents*, Hermès sciences publication (2011).

[30] VAN DER AALST (W.). – *Process Mining : Overview and Opportunities*, ACM Transactions on Management Information Systems (3:2), pp. 7:1-7:17 (2012).

[31] WITTEN (I.H.), FRANK (E.) et HALL (M.). – *Data Mining : Practical Machine Learning Tools and Techniques* (3rd ed.), San Francisco : Morgan Kaufmann (2011).

[32] WU (X.), KUMAR (V.), QUINLAN (J.R.), GHOSH (J.), YANG (Q.), MOTODA (H.), MCLACHLAN (G.J.), NG (A.), LIU (B.), YU (P.S.), ZHOU (Z.-H.), STEINBACH (M.), HAND (D.J.) et STEINBERG (D.). – *Top 10 Algorithms in Data Mining*, Knowledge and Information Systems (14:1), pp. 1-37 (2007).

[33] ZHENG (J.G.) et PELTSVERGER (S.). – *Web Analytics Overview*, In book : Encyclopedia of Information Science and Technology, Third Edition, Publisher : IGI Global, Editors : Mehdi Khosrow-Pour (2015).

[34] KAUVALI (B.), KNOTT (D.) et VAN KUIKEN (S.). – The big data revolution in helthcare : accelerating value and innovation. Mc Kinsey (2013) http://www.mckinsey.com/insights/health_systems/The_big_data_revolution_in_US_health_care.

À lire également dans nos bases

BENARAMA ZITOUNE (F.). – *Analyse automatique d'opinions. États des lieux et perspectives* [H 7 270] (2016).

FIGER (J.P.). – *Coud computing et informatique en nuage*. [H 6 020] (2012).

MELANÇON (G.). – *Visualisation d'informations*. [H 7 417] (2015).

KEMBELLEC (G.), CHEVALIER (M.) et DUDOGNON (D.). – *Systèmes de recommandation*. [H 7 245] (2015).

DELORT (J.Y.). – *Génération automatique de résumés*. [H 7 260] (2007).

Conférences

ICDM – International conference on Data Mining
http://lcdm2016.eurecat.org

TECHNIQUES DE L'INGÉNIEUR

L'expertise technique et scientifique de référence

Techniques de l'Ingénieur vous apporte une information précise et fiable pour l'étude et la réalisation de vos projets.

Actualisées en permanence, les **ressources documentaires** profitent aujourd'hui à plus de **300 000 utilisateurs** et sont la référence pour tout ingénieur, bureau d'études, direction technique et centre de documentation.

Depuis près de 70 ans, **3 500 experts** contribuent quotidiennement à développer, enrichir et mettre à jour cette documentation professionnelle unique en son genre.

L'intégralité de ces ressources représente plus de **9 000 articles**, répartis dans plus de **430 bases documentaires**, accessibles sur internet, en téléchargement PDF, et sur tablette.

4 BONNES RAISONS DE CHOISIR TECHNIQUES DE L'INGÉNIEUR

- Une actualisation permanente du fonds documentaire
- Un comité d'experts scientifiques et techniques reconnus
- Une collection scientifique et technique incontournable sur le marché francophone
- L'espace actualité pour suivre les tendances et innovations de vos secteurs

DES SERVICES ASSOCIÉS À CHAQUE ABONNEMENT

- **Service de questions-réponses** [1][2] **:** interrogez les plus grands spécialistes des domaines couverts par vos bases documentaires. Votre abonnement vous permet en effet de poser des questions techniques ou scientifiques.
- **Les articles Découverte :** un article vous intéresse, mais ne fait pas partie de votre abonnement ? Techniques de l'Ingénieur vous offre la possibilité de l'ajouter.
- **Le Dictionnaire technique multilingue :** 45 000 termes scientifiques et techniques – avec illustrations et légendes – en français, anglais, espagnol, allemand.
- **Les Archives :** vos bases documentaires s'enrichissent et sont mises à jour en ligne en permanence. Les Archives conservent la mémoire de ces évolutions et vous permettent d'accéder aux versions antérieures de vos articles, ainsi qu'à ceux qui traitent des technologies plus anciennes.

Profitez également de l'impression à la demande [1], pour commander une ou plusieurs éditions papier supplémentaires de vos bases documentaires (sur devis).

(1) Disponible pour la France, le Luxembourg, la Belgique, la Suisse et Monaco.
(2) Non disponible pour les établissements scolaires, écoles, universités et autres organismes de formation.

ILS NOUS FONT CONFIANCE :

Sagemcom · EDF · Safran · École Polytechnique · Arkema · AgroParisTech · 3M · Siemens · Rolex · ACS · Dassault Aviation · EADS · L'Oréal · Saint-Gobain · GE · Schneider Electric · Thales · Total

TECHNIQUES
DE L'INGÉNIEUR

Pour disposer d'un panorama complet sur une thématique

DÉCOUVREZ
les offres de packs !

LES + DES OFFRES PACK

- Un large choix de **+ de 60 thématiques** pour des besoins de contenu plus larges
- Des **tarifs préférentiels sur mesure** adaptés à vos besoins

LES UNIVERS DOCUMENTAIRES

- Plus de 430 bases documentaires et plus de 9 000 articles en 14 univers

- Sciences fondamentales
- Environnement - Sécurité
- Énergies
- Technologies de l'information
- Mécanique
- Innovations
- Génie industriel
- Biomédical - Pharma
- Procédés Chimie -Bio - Agro
- Matériaux
- Mesures - Analyses
- Électronique - automatique
- Construction
- Transports

POUR EN SAVOIR PLUS SUR LES OFFRES DE PACKS...

... contactez le service Relation Clientèle
qui se chargera de vous rediriger vers un chargé d'affaires :

Tél : +33 (0)1 53 35 20 20

Email : infos.clients@teching.com
www.techniques-ingenieur.fr

LES AVANTAGES TECHNIQUES DE L'INGÉNIEUR

Le droit d'accès, annuel ou pluriannuel, permet une consultation illimitée des ressources documentaires sélectionnées, ainsi que le téléchargement des versions PDF des articles de référence ou fiches pratiques inclus dans ces ressources. Les droits d'accès sont proposés en monoposte ou multiposte.

- **ACTUALISATION PERMANENTE**
  Mises à jour permanentes, publication de nouveaux articles de références et fiches pratique : un contenu complet sur le sujet qui vous intéresse, des alertes par email.

- **DES SERVICES INCLUS**
  En plus de l'accès aux ressources documentaires, chaque souscription offre un accès privilégié à un ensemble de services.

- **MOBILITÉ** Available on the App Store
  Votre abonnement étant 100 % web, vous pouvez le consulter à tout moment, sur n'importe quel ordinateur ou sur nos versions iPad et Android.

TECHNIQUES DE L'INGÉNIEUR

Pour accompagner vos équipes et projets, CHOISISSEZ les offres de formation et conseil

MONTEZ EN COMPETENCE

- Des formations personnalisées, réalisées au sein de votre établissement et à vos dates
- Un accompagnement à la mise en conformité réglementaire
- Des missions d'audit et de recommandations techniques

LES ENGAGEMENTS TECHNIQUES DE L'INGÉNIEUR

- Un réseau d'experts reconnus pour vous conseiller
- Une veille scientifique et technique pour mieux décider
- Les dernières obligations HSE pour être en règle
- Les clés en management des hommes et des projets pour gagner en efficacité

Consultez l'intégralité des programmes sur le site de Techniques de l'Ingénieur, espaces FORMATION et CONSEIL

www.techniques-ingenieur.fr

RESSOURCES DOCUMENTAIRES | FORMATION | CONSEIL

Nous contacter : Tél. : +33 (0)1 53 35 20 20 – Fax : +33 (0)1 53 26 79 18 – E-mail : infos.clients@teching.com