Big Data

Le cas des systèmes d'information

Josiane Mothe[1,2], Yoann Pitarch[1] , Eric Gaussier[3]

1. Institut de Recherche en Informatique de Toulouse
CNRS, UMR 5505, Université de Toulouse, 118 Route de Narbonne
F-31 062 Toulouse Cedex 9, France
prenom.nom@irit.fr

2. Ecole Supérieure de Professorat et de l'Education
Ecole Interne Université de Toulouse II le Mirail

3. Université Grenoble Alpes / CNRS - Laboratoire d'Informatique de Grenoble
prenom.nom@imag.fr

RESUME. *Nous présentons dans cet article les principaux défis que pose le « big data » aux systèmes d'information, c'est-à-dire aux systèmes en charge du stockage et du traitement des données en vue de prises de décision. Après avoir détaillé deux applications majeures du big data que sont la recherche d'information et l'intelligence économique, nous nous intéressons à la place des données ouvertes et du web dans le big data ainsi qu'à celle que le web occupe dans les sciences et la société. Nous abordons ensuite les méthodes et technologies informatiques déployées pour traiter le big data en mettant l'accent sur la façon dont les données sont stockées, traitées et analysées afin d'en extraire des connaissances. Nous nous intéressons enfin aux défis que pose le big data aux entreprises et aux citoyens, notamment en terme de qualité des données et de préservation de la vie privée.*

ABSTRACT. *In this paper, we present the main challenges "big data" raises to information systems, that is to systems dedicated to the storage and processing of data for decision making purposes. After presenting in detail two major applications of big data (information retrieval and business intelligence), we investigate the role of open data and the web in big bata applications, as well as the role the web plays in science and society. We discuss the methods and computer technologies deployed to address big data challenges, focusing on how the data is stored, processed and analyzed in order to extract knowledge. Finally, we consider the challenges big data raises to companies and citizens, especially in terms of data quality and privacy preserving processes.*

MOTS-CLES : *Systèmes d'information, Big data, recherche d'information, intelligence économique, données ouvertes, Hadoop, NoSQL, fouille de données)*

KEYWORDS: Information systems, big data, information retrieval, business intelligence, open data, Hadoop, NoSQL, data mining

1. Introduction

L'année 2013 est certainement celle qui a rendu le terme « big data » populaire ; les médias se sont emparés du phénomène et de nombreux magasines y ont consacré un numéro spécial, comme (La Recherche, 2013) ou (Pour la Science, 2013). Dans cet article nous souhaitons introduire le *big data* d'une part en termes applicatifs, d'autre part en termes méthodologique, technique et technologique. Le choix a été fait de garder le terme anglophone ; une traduction pourrait être « masse de données ».

L'intérêt grandissant pour les termes « big data » ou « *Hadoop* » doit certainement être relativisé. La fréquence des requêtes sur le moteur *Google* est illustrée à la figure 1 et montre que ces termes connaissent un intérêt grandissant, qui reste cependant limité comparé à celui porté sur les bases de données, ou sur les chiens (cela ne s'invente pas, requête assez stable en termes de fréquence sur les huit dernières années).

La définition technologique du *big data* est maintenant consensuelle (Doug, 1996), (Russom, 2011), (Chen et Zhang, 2014). Elle s'appuie sur les 3V:

- Volume : référence à la quantité d'informations, trop volumineuses pour être acquises / stockées / traitées / analysées / diffusées dans leur entièreté par les techniques actuelles,
- Variété : référence à l'hétérogénéité des formats et de la qualité des informations,
- Vélocité : référence à l'aspect dynamique/temporel des données, qui sont produites en flots continus et/ou sur lesquelles des décisions temps réels doivent être prises.

A ces 3V s'ajoutent des éléments complémentaires comme :

- Valeur : référence à la potentialité des données, en particulier en termes économique,
- Véracité / Validité : référence à la qualité des données et/ou aux problèmes éthiques liés à leur utilisation.

Ces différents aspects sont traités dans la suite.

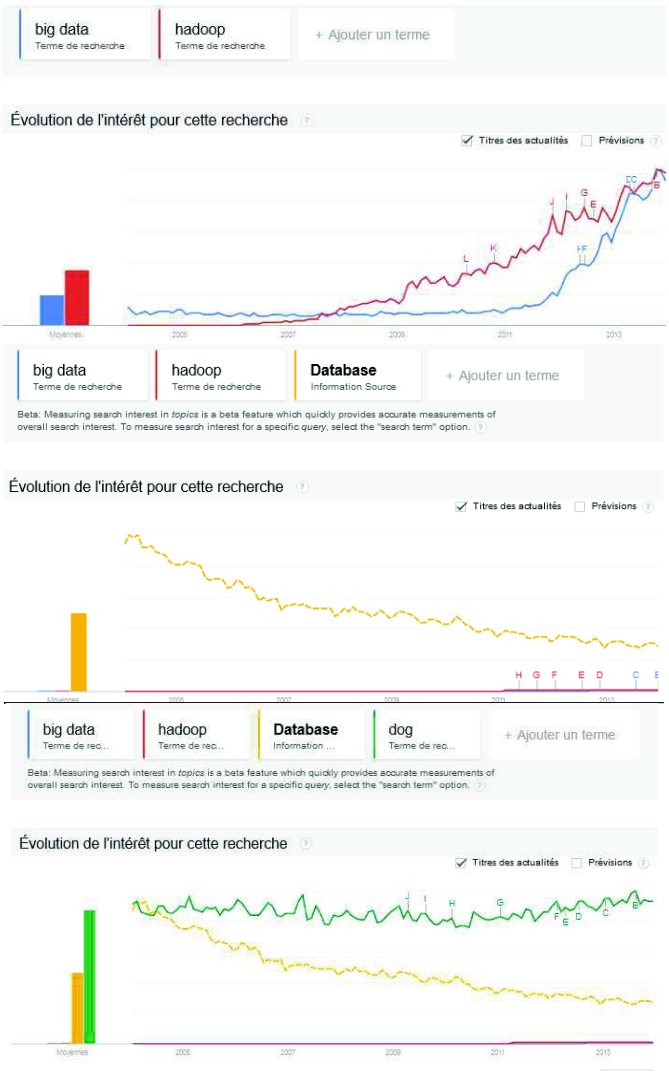

Figure 1 : tendance de la fréquence des requêtes de 2005 à aujourd'hui (par *Google trends*). En bleu « *big data* », en rouge « *hadoop* », en jaune « *database* » et en vert « *dog* ».

Figure 2. Ce que recouvre les 3V du *big data*- extrait de (Russom, 2011)

L'aspect volume est certainement celui qui est le mieux décrit par le terme « big » de l'expression. De nombreux articles et diaporamas regorgent d'exemples relatifs à l'augmentation impressionnante des volumes de données numériques produites dans le monde. Les statistiques sur l'usage d'Internet sont impressionnantes avec 35% de la population mondiale utilisatrice soit pratiquement 2 milliards utilisateurs[1], 1,37 millards de pages web indexées[2], 1,19 millards d'utilisateurs actifs mensuellement sur Facebook, 58 millions de tweets en moyenne envoyés par jour[3] (de nombreuses autres statistiques sur le Web 2.0 sont disponibles[4]). A cela, il convient d'ajouter les données produites par les entreprises qui ne sont pas sur les services Internet. Même s'il existe une controverse sur la déclaration de E. Schmidt concernant l'exactitude des chiffres, personne ne doute que la cadence de production de données s'accélère (on parle de 5 ExaOctets produits tous les deux jours). Le *big data* pourrait alors être défini comme dans le rapport McKinsey « *Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze* » (Manyika et al., 2011). Cette vision remonte au moins à 1997. Dans leur article, Cox et Ellsworth de la NASA distinguent les « *big data objects* » et les « *big data collections* » (Cox et Ellsworth, 1997). Les premiers font référence à des objets qui sont unitairement trop gros pour être traités par le logiciel ou matériel disponible (un même objet peut être gros ou pas, selon les applications ou en fonction du matériel

[1] www.internetworldstats.com/stats.htm
[2] http://www.worldwidewebsize.com/
[3] http://www.statisticbrain.com/twitter-statistics/
[4] http://www.socialbakers.com/

disponible). Les secondes étant constituées de petits et gros objets, souvent hétérogènes (Cox et Ellsworth, 1997).

Comme nous le verrons dans la suite de cet article, il n'existe que peu, voire pas, d'applications mettant en jeu des données possédant toutes ces caractéristiques. Une définition plus juste 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 ou de les traiter par les algorithmes actuels ». Ainsi, l'aspect *a priori* évident du volume du *big data* ne doit pas cacher les deux autres aspects fondamentaux que sont la variété et la vélocité des données.

Comme le soulignent McAfee et Brynjolfsson (2012), pour certaines applications, la vitesse de création des données est plus problématique que leur volume. Par exemple, les données que sont les informations sur la position d'objets mobiles doivent être traitées en temps réel au risque de devenir obsolètes et inutiles. La vélocité est particulièrement importante également dans le cas des données issues des capteurs ou de vidéo-surveillance qui nécessitent un traitement rapide pour une réaction en temps réel. Si la plupart des données *big data* sont des données WORM[5], dans le cas de données de grande vélocité telles que citées précédemment la possibilité de ne pas les stocker ou de les résumer avec précision est à envisager.

La variété des données, comme celle du volume en fait (Kass, 1980), n'est pas nouvelle en soit (Arens *et al.*, 1993). Cependant cette variété se posait jusqu'ici plus en termes de schémas de base de données hétérogènes et des solutions ont été apportées dans ce domaine, en particulier avec les bases de données réparties et fédérées (Zurfluh *et al.*, 1993) (Soutou, 1994). Aujourd'hui, cependant, le *big data* prend la forme de documents, images, vidéos mais également de messages comme les courriels, chats, posts sur les réseaux sociaux, de traces laissées sur les réseaux (téléphone, Internet), de données issues de capteurs ou des CRM[6], des données structurées issues des bases de données relationnelles, etc. autant de données dont les formats et les structures n'ont pas été pensés pour réaliser des liens entre elles.

Associés à cette définition technologique, les aspects économiques sont souvent associés à la Valeur du *big data*. Les prévisions en termes d'impact économique sont importantes. Selon Peter Soundergaard, vice-président sénior de Gartner Inc., d'ici 2015, 4,4 million d'emplois dans les technologies de l'information seront créés pour soutenir le *big data*, dont 1,9 million aux Etats-Unis. e-skills UK et SAS prédisent qu'un tiers des organisations de 100 employés et plus du Royaume-Uni, soit 6 400, implanteront des programmes de fouille de *big data* dans les cinq prochaines années, conduisant à une demande de 69 000 spécialistes (e-Skills UK, 2013). En termes de Valeur, McKinsey Global Institute indique que dans les seuls Etats Unis, il manquerait environ 150 000 personnes avec une expertise en analyse de *big data*. Cet organisme calcule que le système de santé américain pourrait créer 300 milliards de dollars de valeur par an dont deux tiers correspondrait à des réductions de coût d'environ 8%. Selon leur calcul, 100 milliards d'euros pourraient être économisés

[5] Write Once, Read Many
[6] Customer Relationship Management ou outils de gestion de la relation client

sur les dépenses des états européens, sans compter les réductions liées à la réduction des erreurs et des fraudes.

En relation assez directe avec les définitions précédentes, la Véracité des *big data* est un dernier aspect que nous aborderons. La Véracité fait référence à la qualité des données initiales qui reste souvent nécessaire. Elle comprend d'une part les problèmes de valeurs aberrantes ou manquantes (ces problèmes sont souvent résolus par le volume de données), d'autre part la confiance que l'on peut avoir dans les données.. S'il existe des critères permettant de qualifier la qualité des données (Mothe et Sahut, 2011), dans le cas de *big data*, cette vérification de la qualité est rendue difficile sinon impossible par le volume, la variété et la vélocité.

L'importance du *big data* et son potentiel ne font pas de doute, en particulier si l'on regarde les politiques nationales et internationales de soutien. Au niveau national, le Programme d'Investissements d'Avenir dans le domaine du numérique et dans l'action « cœur de filière », consacrera 150 millions d'euros d'aides à la recherche et au développement aux quatre technologies stratégiques : le logiciel embarqué et les objets connectés, la sécurité des systèmes d'information, le calcul intensif et la simulation, et le *cloud computing* et le *big data*. Par ailleurs, *Policy Exchange*, un des majeurs « *think tank* » du Royaume-Uni a identifié le *big data* comme une des 8 technologies majeures[7]. Au niveau Européen, dans le programme H2020, deux appels y sont tout particulièrement dédiés, l'ICT-15-2014 (*Big data and Open Data Innovation and take-up*), doté de 658,5 millions d'euros et l'ICT-16-2014 *(Big data-research*). Le *big data* touche tous les secteurs et les grands défis sociétaux définis par l'Europe tels que l'adaptation au changement climatique, l'énergie propre sûre et efficace, la santé, le défi démographique, la mobilité, les sociétés innovantes, intégrantes et adaptatives, la société de l'information et de la communication, ou la liberté et la sécurité de l'Europe, de ses citoyens et résidents. Ce dernier point fait clairement référence à la notion de Vie privée que les applications de *big data* risquent d'oublier. Comme le souligne (Pentland, 2013), « les traces numériques que chacun de nous laisse derrière lui chaque jour pourraient devenir un cauchemar pour notre vie privée... ou servir à construire un monde plus sain et plus prospère ».

Dans la section 2 de cet article, nous illustrerons les enjeux sociétaux et économiques à travers plusieurs domaines d'application. La recherche d'information sera d'abord présentée. Ce domaine est certainement le moteur majeur du *big data* : ce sont les entreprises du web telles que *Google* qui ont inventé les nouvelles technologies propres au *big data* et à la gestion des données volumineuses. Nous aborderons ensuite l'intelligence économique qui n'est pas un domaine nouveau mais qui voit ses potentialités décupler avec l'approche *big data*. Les données ouvertes et liées (*open* et *linked data*) offrent de nouvelles perspectives et sont le cadre d'innovations en termes applicatifs et seront décrites. Nous terminerons cette section 2 par les applications en sciences, avec des impacts sociétaux forts attendus, en particulier en santé.

[7] http://www.policyexchange.org.uk/

La section 3 sera dédiée à la présentation des grandes classes de technologies sous-jacentes au *big data*. Concernant le stockage, l'apparition des bases de données NoSQL est un des éléments nouveaux associé au *big data* ; à travers plusieurs catégories de bases de données, elles sont supposées répondre à différents types de besoin en s'affranchissant de la rigidité des bases de données structurées mais en sacrifiant aussi la facilité de requêtage au travers du SQL. L'adaptabilité et l'extensibilité propre au NoSQL (Not Only SQL) est dans la même mouvance que l'extensibilité des ressources matérielles qu'offre le nuage (le cloud). Hadoop et plus spécifiquement la distribution des traitements et des données est un autre élément clé du *big data*. Enfin, la section 3 abordera également les facettes relatives au traitement des *big data* pour leur exploitation dans des applications d'extraction de connaissances.

Finalement, la section 4 terminera cet article en abordant quelques enjeux soulevés par le *big data*, au sein des entreprises par la mise en place d'une démarche *big data* et ses potentiels freins, et de façon plus générale sur les questions éthiques liées au *big data*.

2. Enjeux économiques et sociétaux : applications des *big data*

Dans cette section, nous ne prétendons pas dresser un panorama exhaustif des applications *big data* : la potentialité en est trop vaste. Nous nous focalisons plutôt sur les applications qui de par leur tradition ou leurs avancées concrètes sont définitivement des applications *big data*. Pour chacune des grandes applications que nous aborderons, nous tenterons d'en faire ressortir les dimensions V qui leurs sont plus particulièrement associées.

2.1. Recherche d'information

La recherche d'information concerne les méthodes et les logiciels permettant de donner accès à une information répondant à un besoin de l'utilisateur (Amini et Gaussier, 2012) (Mothe, 2000). Ce domaine n'est pas né avec l'Internet mais y a trouvé une mise en application distribuée très grande échelle. Comme le rappellent (Sanderson et Croft, 2012), UNIVAC d'Holmstrom est certainement la première machine capable de rechercher des documents secondaires à partir de sujets (Holmstrom, 1948), mais les premiers travaux d'envergure débutent fin des années 60, début des années 70, en particulier avec G. Salton (Salton, 1965).

D'un point de vue académique, les collections utilisées pour valider ou évaluer les propositions scientifiques sont à cette époque assez réduites, de l'ordre de quelques milliers de documents courts, comme pour les collections CACM (3 024 documents et 4 326 requêtes), CISI (1 460 documents et 3 918 requêtes), Medline (1 033 documents, 5 134 requêtes) ou Cranfield (1 033 documents et 5 134 requêtes) (Cleverdon, 1966). Au début des années 1990, avant l'arrivée réelle du web, la communauté scientifique se structure autour du projet d'évaluation grande échelle TREC (Text Retreival Conference) qui débutera en 1992, sponsorisé par le *National Institute of Standards and Technology* (NIST) et le département de la défense

américain dans le cadre du programme TIPSTER Text. Au-delà de fournir une base commune en termes de collection et de méthodologie d'évaluation, ce programme se donne également comme ambition de démontrer la capacité des systèmes d'opérer dans des environnements de recherche d'information réels (Harman, 1993). Ainsi, la première collection distribuée dans ce programme contient 1,26 Go de textes. Dans le même temps, en 1990, un disque dur *normal* avait une capacité de 40 Mo. Dans son article, D. Harman souligne que pour les 25 participants internationaux, les efforts ont surtout été mis sur les aspects engineering des systèmes pour leur permettre de gérer le nombre important de documents et que peu d'effort a pu être consacré à la mise au point des systèmes et à leur paramétrage (Harman, 1993). C'est en 2007 que le disque de 1 To est disponible (Hitachi GST). En 2009, TREC met à disposition une collection (ClueWeb09) créée par l'université Carnegie Mellon (Language Technologies Institute) avec 1 milliard de pages web dans différentes langues, correspondant à 25 To décompressé (Clarke *et al.*, 2010). C'est la collection de test la plus volumineuse aujourd'hui. Elle sert de base à des expériences relevant de différentes tâches de recherche d'information.

Parallèlement, le WWW se développe. Tim Berners-Lee, informaticien au CERN l'invente en 1990. Après une mise à disposition interne, le premier serveur web est mis en ligne aux Etats-Unis dans l'institut de recherche SLAC (National Accelerator Laboratory) en 1991. Dès 1994, le web comptait 10 000 serveurs, dont 2000 à usage commercial, et 10 millions d'utilisateurs. (http://home.web.cern.ch/fr/about/birth-web). En 1993, les premiers moteurs de recherche sur le web se développent : JumpStation, the World Wide Web Worm, et Repository-Based Software Engineering (RBSE) spider ; ce dernier étant le premier à implanter un ordonnancement des résultats. Ils s'appuient alors sur le titre et l'adresse des pages. Webcrawler, en 1994 est le premier à indexer le contenu des pages. A la même période AltaVista voit le jour ; il servira de base à Yahoo ! Search (Wall, 2006). *Google* quant à lui verra le jour en 1997 (Google, 2013). En 2008, Alpert et Hajaj ont posté «*We knew the web was big*»: en 1998, le premier index *Google* contenait 26 million pages, en 2000 l'index en contenait 1 milliard (Alpert et Hajaj, 2008). D'après le site worldwidewebsize, fin novembre 2013, ce serait entre 15 et 40 milliards de pages qui seraient indexées[8].

Si la recherche d'information est clairement un domaine dans lequel le volume est au cœur des problématiques dès son avènement, il n'est pas le seul aspect qui rend la recherche d'information une application *big data*.

L'aspect variété des informations, s'exprime d'abord au travers des langues de rédaction des documents, plusieurs pouvant cohabiter (Grefenstette, 1998) (Savoy, 2005) (Peters et al., 2012). La variété de taille et de type des objets recherchés et leur évolution est un autre volet de la variété. Les applications actuelles, tout comme les recherches dans le domaine, sont plutôt spécialisées dans un type d'objets recherchés. Chez *Google* par exemple l'offre se décline en *Google Maps, Google Livres, Google Schoolar, Google Images, Google Vidéos*. Les programmes

[8] http://www.worldwidewebsize.com/

internationaux d'évaluation de la recherche d'information se déclinent de la même façon, autour des formats (Vidéo, texte, tweet), des tâches (diversité, filtrage, contextualisation) ou des domaines (brevets, médical). D'autres types de données concernent les traces de connexions laissées lors de l'utilisation de service. Les moteurs de recommandation combinent leur analyse avec celles des contenus. Ainsi, les modèles dits mixtes proposent non seulement des contenus similaires, selon un ou plusieurs critères, au document consulté par l'utilisateur, mais prennent également en compte les consultations des autres utilisateurs qui ont vu ce document (Resnick et Varian, 1997) (Ricci *et al.*, 2011).

La vélocité et la création de valeur est au cœur des problématiques des moteurs commerciaux. De nombreux outils sont d'ailleurs proposés pour analyser les temps de réponse des services web. Lorsque l'on sait que *Google* répond à plus de 100 milliards de recherche par mois, on comprend l'ampleur du défi. Les principes de cache (pré-calcul des premières réponses en fonction des requêtes) sont une des techniques les plus utilisées pour accélérer les temps de réponse (Macdonald *et al.*, 2012). L'autre défi relatif à la vélocité en recherche d'information concerne l'intégration de nouveaux documents dans les index. En 2012, 300 millions de nouvelles photos ont été ajoutées chaque jour dans le monde sur Facebook, plus de 85 000 posts par mois pour le seul Brésil et 51 millions nouveaux sites web auraient été ajoutés[9].

Enfin, la validité de l'information n'est souvent que très peu prise en compte. Wikipédia a fait une tentative en marquant les articles ne contenant pas suffisamment de références ; les moteurs de recherche intègrent très partiellement cet aspect en favorisant certains domaines (wikipédia le fut un temps, les réseaux sociaux le sont maintenant) ou certaines pages (en particulier des éléments publicitaires qui font partie du modèle économique des moteurs)[10].

Il est clair que l'aspect volume est un aspect qui est au cœur des systèmes de recherche d'information ; le domaine a une histoire face à ce défi et a su y apporter des solutions. *BigTable* (Chang *et al.*, 2006) par exemple, le précurseur de *HBase* implanté en particulier dans Hadoop (cf section 3.2), a été développé chez *Google*. En revanche, l'aspect variété est mal traité et un des grands défis du domaine est de proposer des modèles qui vont au-delà de listes ordonnées de documents, fusionnant et synthétisant des informations structurées et non structurées (Allan et al., 2012).

2.2. Business intelligence

Comme pour la recherche d'information, l'informatique décisionnelle et la veille stratégique ont une longue histoire mais la mise à disposition de masses de données, la variété des sources dont elles proviennent ainsi que leur disponibilité et les

[9] http://royal.pingdom.com/2013/01/16/internet-2012-in-numbers/ consulté le 11 déc. 2013

[10] Cet aspect reflète une constatation de l'auteure, mais ne s'appuie sur aucune étude ou référence.

capacités de stockage et de calcul actuelles ont rendu de nouvelles applications possibles.

Chen et al. (2012) définissent le *business intelligence and analytics* comme les techniques, technologies, systèmes, pratiques, méthodologies et applications qui analysent les données critiques de l'entreprise afin de mieux l'aider à comprendre ses affaires et son marché et à prendre les bonnes décisions au bon moment.

Dans le cadre de l'intelligence économique, stratégique ou politique, il parait difficile de ne pas citer la fouille à grande échelle réalisée par l'Agence nationale de la sécurité américaine (NSA) dans le programme Prism. Ce que l'on peut qualifier de scandale a été dévoilé par Edward Snowden. La NSA aurait ainsi eu accès aux bases de données des acteurs majeurs d'Internet (dont *Google*, *Facebook* et *Microsoft*) pour recherche des documents, des mails ou des chats mais également via les câbles sous-marins (Untersinger, 2013). L'écoute aurait porté en particulier sur des dirigeants européens et d'Amérique latine (téléphone, courriels). Le *patriot act*, loi anti-terroriste américaine, permet également aux services de sécurité d'accéder à des données numériques des particuliers et des entreprises sans autorisation préalable et sans information aux utilisateurs[11]. Les Etats-Unis et la NSA se dotent d'un centre de données à *Bluffdale* (*Intelligence Community Comprehensive National Cybersecurity Initiative Data Center*) pour un montant estimé à 1,7 milliards de dollars, avec une capacité de stockage de 20 To par minute pour une consommation de 65 MW (Carroll, 2013); certains parlent de 10 exa octets voire 1 yotta octet de capacité de stockage total L'intelligence politique est certainement en lien étroit avec l'intelligence économique.

Les brevets jouent un rôle crucial dans l'économie et dans l'évolution technologique. Par exemple, *Thomson Scientific* met à disposition *Derwent Worlds Patents Index* ® qui couvre plus de 14 millions d'inventions (The Thomson Corporation, 2007) et il y a environ 2 million de demandes de dépôt de brevet dans le monde chaque année selon *WIPO* (2011). Les brevets sont donc un terrain important pour l'intelligence économique. La recherche documentaire autour des brevets permet de réaliser une étude du domaine considéré, en y intégrant ses différentes facettes. Ce type de recherche est crucial en recherche, pour les entreprises et les gouvernements, pour connaitre les principales activités, les concurrents et les marchés (Dkaki *et al.*, 1997) (Kim *et al.*, 2012) ou pour détecter les tendances émergentes (Kontostathis *et al.*, 2003) (El Haddadi *et al.*, 2012). Les brevets sont pour partie structurés, mais la majeure partie de l'information ne l'est pas. Leur contenu est varié : à la partie textuelle s'ajoute souvent des schémas explicatifs relatifs à l'invention. Les techniques d'exploration peuvent alors être utilisées. Il est par exemple possible de créer un réseau de brevets comme dans (Tang *et al.*, 2012) où les différents types d'éléments (compagnies, inventeurs et contenus techniques) sont représentés dans un réseau prenant en compte l'évolution temporelle. La veille concurrentielle est également une activité fondamentale pour les entreprises innovantes. El Haddadi *et al.* (2011) proposent ainsi des scénarii

[11] http://fr.wikipedia.org/wiki/USA_PATRIOT_Act ; consulté décembre 2013.

d'utilisation de méthodes d'analyse de données pour détecter les signaux faibles et pour analyser les évolutions et les tendances dans un domaine. Upson *et al.* (2012) s'intéressent au lien entre l'analyse de la concurrence et l'implantation d'une compagnie dans un secteur dans lequel elle n'était pas encore positionnée.

Ce type d'application est également lié au marketing pour lequel l'analyse de tendances peut permettre d'optimiser les chances de succès lors de l'introduction de nouveaux produits. Le marketing relationnel (centré client) vient remplacer le marketing transactionnel (centré produit). Il s'agit de fidéliser le client en entretenant une relation privilégiée et personnalisée, plutôt que se concentrer sur la réalisation d'une vente. Ce changement est en particulier possible par l'accès à des informations issues des CRM de l'entreprise mais également des réseaux sociaux. Les communautés permettent par exemple à un ensemble de personnes de discuter de sujets d'intérêt commun (McWilliam, 2000) et de nombreuses compagnies proposent des lieux de communication *customers-to-customers* en lien avec leurs produits ou services. Dans son livre blanc, Gigya identifie les réseaux sociaux comme des sources pour augmenter la notoriété des produits ou des compagnies, mais également des lieux sur lesquels les entreprises peuvent agir pour apporter de la valeur tout en maintenant la transparence des échanges et des usages (Gigya, 2013). En parallèle, la stratégie d'achat d'espaces publicitaires se voit modifiée en particulier autour des plateformes d'achat de ces espaces en temps réel (*Ad Exchange*).

Ce sont donc des flots d'information qui doivent être traités plutôt que des sources plus figées comme cela était le cas il y a quelques années (Davenport *et al.*, 2012). La surveillance en temps réel est rendue possible. L'analyse associée peut ainsi permettre la détection de l'évolution de l'opinion des consommateurs sur des entreprises, des marques ou des produits, ainsi que de nouveaux besoins. Il peut s'agir d'analyser l'image numérique ou de détecter les sources ou les individus les plus influents (par exemple les sources qui sont *re-tweetées*) afin de faire évoluer cette image numérique et de répondre rapidement lorsqu'un risque de dégradation d'image apparaît (Barbieri *et al.*, 2013). De façon similaire, l'analyse de sources multiples peut permettre de détecter rapidement un problème (technique sur un produit, fuite dans un circuit par exemple) ou une défaillance (sur une chaine de production) dès qu'il surgit. L'analyse porte alors sur des données collectées sur des capteurs ou des échanges entre individus.

Il s'agit donc d'aider la prise de décisions et la réaction rapide voire en temps réel. Ainsi Davenport et al. (2012) indiquent que compte tenu de l'évolution *big data*, l'architecture des applications va changer. Plutôt que de proposer des boites noires associées à des entrepôts de données, les nouvelles applications devront selon eux évoluer vers des éco-systèmes dans lesquels les services internes et externes devront s'échanger des informations en continu pour générer des éléments utiles pour l'organisation dans ses prises de décision.

Les informations externes proviennent de plus en plus des données ouvertes et du web de données. Ces données font l'objet de la section suivante.

2.3 Données ouvertes et Web des données

Le mouvement des données ouvertes et celui des données liées ont largement contribué au succès des *big data* en mettant à la disposition de la communauté une multitude de données hétérogènes au fort contenu sémantique. Dans la suite de ces sections, nous décrivons ces deux mouvements.

2.3.1. Données ouvertes

L'ouverture des données peut être vue comme une philosophie dont l'objectif est la mise à disposition de données numériques, publiques ou privées, à tous et sans restriction sur le droit d'accès ou de réutilisation.

Le monde universitaire est considéré comme pionnier dans l'ouverture des données. En effet, le partage du savoir et de la connaissance est apparu très tôt comme une nécessité. A cet effet, la déclaration internationale sur le libre accès, rédigée à Budapest en 2001, fut un véritable appel à la communauté scientifique pour la mise à disposition gratuite des résultats de recherche. A titre d'exemple, le projet sur le séquençage du génome humain a été permis par l'«*Open Data Consortium*[12] » afin de permettre aux pays en voie de développement ou aux universités (n'ayant pas toujours accès à certaines revues scientifiques) de pouvoir bénéficier de cette connaissance et ainsi l'exploiter dans leurs recherches.

D'autres pionniers de l'ouverture des données sont les gouvernements. On parle alors de données gouvernementales ouvertes. Sur ce plan, le mouvement des données ouvertes a été initié par le président américain, Barack Obama, dans le mémorandum sur la transparence et le gouvernement ouvert[13] (janvier 2009). L'idée de base était d'établir une coopération moderne entre les politiciens, les administrations publiques, les entreprises et les citoyens pour garantir une complète transparence et ainsi favoriser la démocratie et la collaboration entre les acteurs sociaux. De nombreux pays ont emboité le pas. Ainsi, l'«*Open Government Partnership*» a été lancé en septembre 2011, et qui comptera 62 pays d'ici avril 2014. L'objectif de ce partenariat est de fournir une plateforme internationale pour permettre aux gouvernements membres de travailler ensemble à la diffusion de leurs données.

L'écosystème lié aux données ouvertes est naissant. En tant que tel, il ne peut raisonnablement pas être complètement structuré ni même rentable. La mise en place d'un modèle économique viable représente d'ailleurs un enjeu de taille que les acteurs des données ouvertes doivent considérer pour rendre ce mouvement durable. Plus précisément, deux grandes natures de modèles semblent se dessiner : le modèle des producteurs de données et celui des consommateurs de données. Côté producteurs, la collecte et la mise à disposition de données fraîches, structurées et bien documentées a un coût non négligeable. Si les défenseurs de l'*open source* arguent que ce coût est déjà supporté par les impôts et prônent donc pour une

[12] http://www.opendataconsortium.org
[13] http://www.whitehouse.gov/the-press-office/transparency-and-open-government

gratuité totale des données ouvertes, certaines entreprises ou institutions (par exemple le Grand Lyon[14]) ont fait le choix de monétiser certains jeux de données. Côté consommateurs de données, les attentes en termes de retombées économiques sont immenses. En effet, la Commission européenne a commandité en 2006 une étude intitulée "Measuring European Public Sector Information Resources" (MEPSIR)[15] pour évaluer le potentiel du marché des données ouvertes. Ce potentiel est estimé à 30 milliards d'euros même si les auteurs nuancent grandement leur propos en évoquant les limites de l'utilisation actuelle des données (utilisation de données à faible voire nul pouvoir de monétisation, nécessité d'ancrer un service gratuit dans les habitudes des utilisateurs pour ensuite le monétiser). Enfin, il est important de remarquer que même côté utilisateurs de données, plusieurs modèles économiques devront perdurer à terme. Typiquement, certains services ne génèreront probablement jamais d'argent et ont plutôt pour vocation de faciliter la vie quotidienne des utilisateurs finaux (par exemple les services de consultations des horaires des transports en commun). A contrario, d'autres ont des ambitions financières bien plus conséquentes (par exemple, IBM avec leur "TheSmarterCity"[16]).

En 2010, la fondation Sunlight a formulé une liste de 10 critères pour qu'une donnée soit considérée comme ouverte[17]. Une donnée ouverte doit être : (1) complète, (2) primaire, (3) opportune, (4) accessible, (5) exploitable, (6) non-discriminatoire, (7) non-propriétaire, (8) libre de droits, (9) permanente et (10) gratuite. La même année, Tim Berners-Lee a proposé une échelle pour indiquer la qualité d'une donnée ouverte[18]. Celle-ci va d'une à cinq étoiles : (*) correspond au données non-filtrées et sans contrainte sur le format, (**) désigne les données structurées, (***) est associé aux données libres d'utilisation (aussi bien sur le plan juridique que sur le plan du format de diffusion), (****) correspond aux données sont accessibles via une URL et (*****) s'applique aux données liées à d'autres données.

Ce dernier point fait apparaître un point clé dans l'exploitation et la contextualisation des données : celles-ci se doivent d'être connectées pour que l'on puisse en tirer le maximum de sémantique. Le mouvement du web des données décrit dans la suite de cette section remplit cet objectif.

2.3.2. Web des données

Le web des données (ou *Linked Data* en anglais) peut être défini comme l'utilisation conjointe de RDF (*Resource Description Framework*) et de HTTP (*Hypertext Transfer Protocol*) pour publier sur le web des données structurées provenant de différentes sources et les connecter entre elles. Les principes régissant

[14] http://smartdata.grandlyon.com
[15] http://www.epsiplatform.eu/content/mepsir-measuring-european-public-sector-resources-report
[16] http://www-03.ibm.com/innovation/us/thesmartercity/
[17] http://sunlightfoundation.com/opendataguidelines/
[18] http://5stardata.info

ces données ont été formulés pour la première fois dans Berners-Lee (2006). Ce guide a ensuite été étendu dans divers documents techniques tels que Bizer *et al.* (2007) ou Sauermann *et al.* (2011). Ces documents regroupent un ensemble de bonnes pratiques dédiées aux fournisseurs de données afin de rendre leurs données parfaitement interopérables.

Similairement aux navigateurs qui permettent d'évoluer efficacement au sein du « *web* des documents » par le biais des URLs, des navigateurs particuliers existent pour les « *linked data*». Ceux-ci exploitent les relations RDF qui permettent de spécifier que deux ressources sont connectées entre elles. Par exemple, un lien RDF entre deux ressources représentant des individus peut vouloir dire que ces personnes se connaissent. Dans un contexte académique, un lien entre une ressource associée à un chercheur et une autre ressource associée à une publication scientifique peut matérialiser la relation « être auteur de ».

Parmi les preuves les plus flagrantes que le *web* des données connaît un essor significatif, on peut citer le projet « *Linking Open Data*[19] ». Crée en 2007, son objectif est de regrouper les efforts de la communauté pour (1) identifier les jeux de données libres de droits, (2) les re-publier au format RDF et (3) connecter les jeux de données entre eux. Ce projet, soutenu par le « *W3C Semantic Web Education and Outreach Working Group* » compte actuellement 295 jeux de données, rassemblant plus de 31×10^9 triplets RDF et 503×10^6 liens. Un point particulièrement intéressant concernant ce projet est qu'il couvre des thématiques très diverses, allant des médias aux données scientifiques en passant par la géographie, permettant ainsi la création d'applications innovantes.

2.4. Science et société

Dans différents domaines scientifiques le *big data* est synonyme de mise à disposition de masse énorme de données collectées à travers des expériences ou des capteurs, pour favoriser la résolution de problèmes de recherche via l'intelligence collective. Les grands défis sociétaux tels que la santé et l'éducation trouvent leur place dans le *big data*. Les données sont souvent disponibles, mais de nouveaux moyens de mise à disposition et d'analyse doivent être mis en place comme le souligne dans son éditorial le journal *Nature* en 2008 : « *Above all, data on today's scales require scientific and computational intelligence. Google may now have its critics, but no one can deny its impact, which ultimately stems from the cleverness of its informatics. The future of science depends in part on such cleverness again being applied to data for their own sake, complementing scientific hypotheses as a basis for exploring today's information cornucopia.* » (Nature, 2008). Ceci est d'autant plus vrai que les mêmes données peuvent être utilisées par différentes communautés ou pour différents objets de recherche.

La NASA a par exemple mis à disposition de la communauté scientifique et éducative des données issues de satellites sur le climat et la Terre (température,

[19] http://linkeddata.org

précipitation, couverture forestière)[20]. Ce projet de mise à disposition de données s'intègre au programme national américain d'*open data* et ouvre des possibilités en particulier dans le domaine de l'étude du changement climatique. Les volumes de données produits par les instruments de mesure sont colossaux : en astronomie le télescope LSST (*Large Synopsis Survey Telescope*) installé au Chili enregistre 30 Teraoctets d'images par jour. Le télescope européen quant à lui produira ½ Petaoctets par jour en 2018. Le LHC (*Large Hadron Collider*), l'accélérateur de particule, va engendrer 60 Teraoctets de données par jour.

Dans ces contextes, l'aspect vélocité est un défi dans la mesure où les données non stockées sont perdues et ne pourront pas être reproduites. Il s'agit donc d'être capable d'une part de recevoir les données volumineuses au fur et à mesure qu'elles sont produites ou collectées mais également d'être capable soit de les stocker, soit de les traiter en flot continu. Les solutions pensées actuellement sont entre les deux : il s'agit de stocker des données prêtes pour la science (*science ready data*). Ainsi, les données initiales sont pré-traitées (nettoyage, agrégation) et mise à des formats utilisables. Il s'agit d'un compromis entre traitement et stockage dans la mesure où les traitements qui pourront être réalisés dans le futur sont pour la plupart inconnus.

En génomique, les séquençages de l'ADN et des protéines issus de différentes sources sont collectées (GenBank, RefSeq, …) et mis à disposition de la recherche[21]. L'accès libre aux données est à l'origine du succès de la bio-informatique. Le volume des données soumises à ces banques de données double environ tous les 18 mois mais le temps de doublement semble s'allonger. Perrière (2013) indique que ce phénomène est potentiellement lié au fait que les responsables des centres de données sont de moins en moins à même de supporter les frais de gestion, mais qu'il peut également être lié au problème de transfert de données, difficile via le réseau. Dans le domaine médical le *big data* ouvre des possibilités vers des applications pour la société. La combinaison de données variées, provenant à la fois de dossiers patients, de la biologie, de la génomique, de données sur l'environnement, sur les modes de vie permet d'imaginer une meilleure compréhension des maladies ou du vieillissement et de nouvelles approches pour améliorer la santé.

Dans d'autres domaines scientifiques, les données ne sont pas collectées à partir d'un instrument de mesure mais produites par simulation. Dans ce cas les données étant produites par programme, elles peuvent être à nouveau produites par les mêmes algorithmes et ne nécessite donc pas d'être stockées; le problème de leur stockage est donc moins crucial même s'il existe dans le cas où la production de données est coûteuse.

3. Méthodes et technologies du *big data*

Dans cette section, nous décrivons les changements nécessaires qu'il a fallu apporter aux systèmes d'information pour assurer leur passage à l'âge du *big data*.

[20] http://data.nasa.gov
[21] http://www.ncbi.nlm.nih.gov

Nous abordons dans un premier temps les solutions mises en œuvre pour stocker ces données, puis détaillons les évolutions dans l'analyse des données.

3.1. Stockage

Pour pouvoir interroger et analyser des *big data*, il est d'abord nécessaire de pouvoir les stocker efficacement. Ainsi, cette section est consacrée à l'analyse des méthodes et technologies déployées pour le stockage des *big data*. Dans un premier temps, nous aborderons l'aspect « architecture » en évoquant l'informatique dans les nuages (en anglais *cloud database* et *cloud computing*), ses motivations et son principe de fonctionnement. Enfin, nous adopterons un point de vue « système de gestion bases de données ». Nous aborderons les désormais populaires bases de données NoSQL ainsi que les bases NewSQL dont l'émergence est plus récente.

3.1.1. Gérer le volume des données dans les nuages

Depuis une dizaine d'années, l'informatique dans les nuages (*cloud computing*) connaît un succès qui ne semble pas prêt de décroitre. Si l'on peut penser que ce concept est relativement récent, il n'en est rien. En effet, ses concepts fondamentaux datent des années 1950 quand des mainframes colossaux firent leur apparition dans les universités et les entreprises (Velete *et al.*, 2009). De simples terminaux étaient alors suffisants pour se connecter à ces serveurs et ainsi utiliser des services, non plus stockés localement, mais à distance. Les bases du *cloud computing* étaient nées. Schématiquement, un nuage est un ensemble de ressources, physiques et logicielles, qui fournissent des services à un individu connecté à ces ressources par le biais d'Internet. Le *cloud computing* se distingue des mainframes précédemment évoqués par plusieurs caractéristiques dont les principales sont les suivantes. D'abord, l'accès aux ressources dans le contexte du *cloud computing* se fait essentiellement par Internet alors qu'il s'agit d'un réseau local pour les mainframes. Enfin, les modèles économiques en jeu divergent. L'accès à la ressource est fréquemment en accès libre dans le cadre des mainframes alors qu'elle est un bien monnayable dans le cloud computing, par exemple avec les services de type « Software As A Service ». Les principales caractéristiques d'un nuage sont (Armbrust *et al.,* 2010) :

- L'accès libre et « à la demande » aux ressources ;
- La facilité d'accéder aux services par le biais d'Internet;
- La mutualisation de ressources (matérielles et physiques) ;
- Le paiement à l'usage.

Dans le contexte des *big data*, le stockage de données dans les nuages fait alors particulièrement sens. En effet, cette architecture est prévue pour passer à l'échelle horizontalement grâce notamment au fait de pouvoir facilement mutualiser des ressources hétérogènes. Ainsi, au fur et à mesure que les besoins de stockages grandissent, de nouveaux serveurs sont déployés dans les nuages de façon transparente pour l'utilisateur. Cette architecture est donc particulièrement adaptée pour traiter l'aspect volume des *big data* (Agrawal *et al.*, 2011).

En considérant les points précédents, il n'est pas possible d'affirmer si les aspects variété et vélocité peuvent être satisfaits par l'informatique dans les nuages. Pour ce faire, nous devons considérer un niveau de granularité un peu plus fin : les systèmes de gestion de bases de données.

3.1.2. Systèmes de gestion de bases de données : quelles solutions pour les big data ?

Dans un environnement aussi distribué que le cloud computing, le maintien des propriétés ACID (Atomicité, Cohérence, Isolation et Durabilité) des systèmes de gestion de bases de données relationnels apparaît difficile à garantir (Kossmann *et al., 2010)* et il est d'ailleurs légitime de s'interroger sur la pertinence de les maintenir systématiquement, par exemple, pour gérer des mises à jour de statuts dans un réseau social ou une liste de produits. Typiquement, assurer efficacement la cohérence des données lorsque celles-ci sont stockées sur des dizaines, voire des centaines, de nœuds différents est très difficile en particulier pour prendre en compte les cas où un nœud est hors-service. Dans ce contexte, trois propriétés sont généralement demandées aux systèmes de gestion de bases de données (Pritchett, 2008). En opposition aux propriétés ACID, celles-ci sont regroupées sous l'acronyme BASE (*Basically Available*, *Soft state*, *Eventual consistency*) :

- La disponibilité : sur le Web, il est crucial que les sites soient disponibles la grande majorité du temps. Dans un environnement largement distribué, l'erreur, par exemple, un nœud défaillant, ne saurait plus être considérée comme l'exception mais comme la règle. Des mécanismes de gestion de l'erreur doivent alors être mis en place (Niranjan Mysore *et al.,* 2010) ;
- La tolérance à la partition : il s'agit d'être certains que les opérations de lecture et écriture sur des données répliquées seront redirigées sur les nœuds appropriés ;
- La cohérence finale: cette propriété garantit que si aucune mise à jour n'est apportée sur les données pendant un certain temps, alors tous les nœuds du système seront consistants une fois que les mises à jour auront été propagées.

D'après le théorème de Brewer (2010), aucun système ne peut assurer à la fois la cohérence, la disponibilité et la possibilité d'être partitionné. En outre, dans un système à grande échelle, le partitionnement des données est indispensable. Ainsi, un système de gestion de bases de données devra faire le choix entre la cohérence et la disponibilité. Comme nous l'avons vu, les bases de données relationnelles préfèrent privilégier la cohérence. Dans la mesure où la plupart des applications Web font le choix de la disponibilité, il a fallu mettre au point un nouveau modèle de stockage des données. En outre, le modèle relationnel impose une certaine rigidité dans la structure des données manipulées par l'intermédiaire de la définition de schémas. Cette rigidité peut s'avérer limitante car de plus en plus de données sont

peu ou pas structurées. C'est dans ce contexte favorable à l'évolution du modèle relationnel que sont apparues les bases de données NoSQL.

Les bases de données NoSQL apportent une réponse à la question « Comment une base de données peut-elle massivement passer à l'échelle et gérer des données à la structure flexible ? ». Les principales caractéristiques des bases de données NoSQL sont :

- L'abandon du modèle relationnel ;
- La non-nécessité de définir un schéma fixe sur les données ;
- La possibilité de passer à l'échelle horizontalement grâce à des mécanismes de gestion de l'erreur et de réplication ;
- Les solutions proposées sont pour la plupart *open-source* et activement soutenues par la communauté des logiciels libres.

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[22]. Cependant, ces approches peuvent globalement être regroupées en quatre catégories que nous décrivons succinctement :

- Les bases orientées « clé/valeur » : les données sont stockées sous la forme de grandes tables de hashage distribuées. Ce mécanisme peut alors facilement passer à l'échelle. Les bases NoSQL populaires rentrant dans cette catégories sont : Memcached, Amazon's Dynamo, Project Voldemort ;
- Les bases orientées « documents » : dans ces bases, les données sont stockées à l'intérieur de documents. Un document peut être schématiquement défini comme un ensemble de collections, de mapping et de données scalaires. Finalement, 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. Les bases NoSQL populaires rentrant dans cette catégorie sont : MongoDB, CouchDB, RavenDB ;
- Les bases orientées « colonnes» : dans ces bases, les attributs sont regroupés en famille de colonnes. Ainsi, deux attributs qui sont fréquemment interrogées ensemble seront stockés au sein d'une même famille de colonnes. Les bases NoSQL populaires rentrant dans cette catégories sont : Cassandra, HBase ;
- Les bases orientées « graphe » : elles permettent de représenter les données sous la forme de graphes. Les entités sont alors les nœuds du graphe et les relations que partagent les entités sont alors des arcs qui

[22] Nous invitons les lecteurs intéressés par un aperçu complet des solutions NoSQL existantes à considérer (Cattell, 2011).

relient ces entités. Les bases NoSQL populaires rentrant dans cette catégorie sont : Neo4J, Infinite Graph, OrientDB.

Malgré les indéniables avantages qu'apportent les bases de données pour des applications *big data*, essentiellement Web, quelques arguments justifient qu'elles ne soient pas adoptées dans certaines situations (Leavitt, 2010) :

- Complexité des traitements : dans la mesure où il n'est pas possible d'utiliser SQL pour interroger ces bases, les requêtes complexes doivent être écrites dans un langage de programmation tiers. Si cette tâche peut être aisée pour des requêtes simples, il en est tout autre dans le cas de requêtes complexes ;
- Non support des propriétés ACID : le fait de relâcher la cohérence permet certes un grain de performances et un passage à l'échelle facilité mais peut être critique pour certaines applications. Typiquement, les applications faisant intervenir des transactions financières, comme pour les sites de e-commerces, requièrent que fiabilité et cohérence soient garanties. L'ajout de ces caractéristiques à une base de données NoSQL impose alors le développement d'une couche logicielle supplémentaire de la part du concepteur de l'application ;
- Une technologie peu familière et une communauté encore réduite : les solutions NoSQL étant pour la plupart *open-source*, très peu de services de support client existent. Ceci est clairement un frein à l'adoption de cette technologie par les entreprises, en outre peu familières avec ces bases.

Pour pallier ces limitations et « réconcilier » le monde SQL et le monde NoSQL, une nouvelle architecture de stockage des données est apparue récemment, les bases NewSQL. En effet, ces systèmes permettent une interrogation des données via SQL tout en garantissant des performances et un passage à l'échelle similaires aux bases de données NoSQL. En outre, ces solutions conservent les propriétés ACID. Parmi les systèmes NewSQL, on peut citer Clustrix[23], NuoDB[24], VoltDB (2010) ou encore F1, récemment proposé par Google (Shute *et al.*, 2013).

3.2. Traitements et données en flot ou distribuées

Le volume des données considérées dans le *big data* nécessite de distribuer à la fois ces données ainsi que les traitements qui en sont faits. La vélocité, quant à elle, nécessite de traiter des données à la volée, sans délai. Ces deux aspects, traitement

[23] http://www.clustrix.com/

[24] http://www.nuodb.com

en flot et traitement distribué, ne sont bien sûr pas exclusifs et peuvent être combinés lorsque les flots de données sont importants[25].

Le traitement distribué de données (on parle également de parallélisation) n'est pas né avec le *big data*, tout comme la majorité des traitements mis en œuvre dans le *big data*; la problématique *big data* impose cependant d'adapter la plupart de ces traitements de façon à les déployer sur les masses de données auxquelles nous sommes maintenant confrontés. L'exemple prototypique de traitement distribué proposé pour le *big data* est certainement l'approche *MapReduce* qui comporte deux étapes principales :

1. L'étape *map* dans laquelle un ensemble de données, découpé en sous-parties, est distribué sur des machines différentes qui effectuent le même traitement sur chacune des sous-parties ;
2. L'étape *reduce* qui consiste à agréger les résultats obtenus par les différentes machines impliquées dans la résolution du problème. La solution sur l'ensemble des données est donc construite à partir des solutions obtenues sur chaque sous-partie.

Cette approche, popularisée par Google, est à la base de plusieurs algorithmes d'accès à l'information dans les grandes masses de données, en particulier dans le web, et a fait l'objet d'une implantation *open-source* distribuée par *Apache* et connue sous le nom de *Hadoop*[26].

Sur la base d'*Hadoop*, et toujours distribué par *Apache*, s'est développé le projet *Mahout*, qui fournit des versions distribuées de plusieurs algorithmes standard 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) en classes ou encore des algorithmes de catégorisation tels que les forêts aléatoires[27]. De façon à être le plus général et complet possible, le projet *Mahout* accueille désormais les versions distribuées d'algorithmes d'apprentissage et de fouille de données même si ceux-ci ne reposent pas sur *Hadoop*. Un autre système construit sur *Hadoop* est *Hive*[28], qui permet d'écrire des requêtes HQL (*Hive Query Language*, proche du standard SQL) qui seront exécutées sur un cluster *Hadoop*. Enfin, nous pouvons également citer *Cassandra*[29], système de gestion de bases de données distribuées, qui intègre aussi *Hadoop*.

La mise en œuvre des algorithmes mentionnés ci-dessus nécessite d'une part de les revisiter pour en proposer une version distribuée, et d'autre part des

[25] On trouve dans la littérature les deux termes « flots » et « flux » de données, que nous considérons comme synonymes.
[26] hadoop.apache.org
[27] Nous renvoyons le lecteur intéressé au site mahout.apache.org
[28] hive.apache.org
[29] cassandra.apache.org

environnements matériels permettant de les exécuter en mode distribué. De tels environnements existent à différents niveaux, représentés par exemple par les machines multi-cœurs ou les grilles de calcul. Le projet DEUS[30], dans lequel a été réalisée la première simulation de l'évolution de l'univers du *Big Bang* à nos jours, a ainsi utilisé un supercalculateur, mis à disposition par le GENCI (Grand Equipement National de Calcul Intensif), qui comporte plus de 92 000 cœurs de calcul et qui est capable de réaliser de 2 millions de milliards d'opérations à la seconde (2 PFlop/s). Comme on le voit, les développements matériels et algorithmiques sont tous deux nécessaires au traitement distribué des données dans le *big data*.

Les limitations actuelles à la distribution des algorithmes d'analyse de données (pris ici au sens large, incluant la fouille, l'apprentissage, l'aide à la décision, la visualisation) résident bien sûr dans les deux aspects (difficulté algorithmique, architecture matérielle et langage associé) que nous venons de mentionner. Ainsi, tous les algorithmes 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. Une autre limitation ici est liée au manque de langages standardisés qui devraient permettre à différents développeurs d'utiliser facilement les implantations parallèles existantes et d'en proposer de nouvelles. *Hadoop*, *Mahout* et *Hive* sont des propositions qui vont dans ce sens, mais elles nécessitent d'être étendues.

Le *big data* ne concerne toutefois pas que les données stockées ; comme nous l'avons souligné plus haut, il touche aussi les données échangées et émises en continu, comme les données en flots sur les médias en ligne ou même les relevés astronomiques. La physique est en fait un gros pourvoyeur de données en ligne, qu'il importe de traiter de manière efficace. Les grands défis que posent les flots de données résident dans le fait que les algorithmes nécessaires pour les traiter ne dispose que d'un espace mémoire réduit (c'est-à-dire qu'il n'est possible de stocker qu'une faible proportion des données reçues) et d'un temps limité pour effectuer les traitements voulus (en particulier, on estime souvent que ces algorithmes ne peuvent effectuer qu'une seule passe sur les données). Avec ces contraintes, calculer des estimateurs simples sur les données tels que les quantiles, qui permettent de donner une estimation empirique de la probabilité d'un événement, peut se révéler compliqué (Cormode *et al.*, 2005). Cette complexité augmente si l'on cherche des résumés des données plus riches ou si les données possèdent une structure complexe, sous forme de graphes par exemple. Les données d'interaction sur les réseaux sociaux sont typiquement représentées par des graphes, et le déploiement d'algorithmes de prédiction de diffusion, par exemple, dans de tels graphes pose de sérieux problèmes algorithmiques.

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,

[30] http://www.deus-consortium.org/

supervisée ou non, et de prédiction sur les séries temporelles[31]. Elle a aussi remis en avant les travaux sur la construction incrémentale de modèles, ainsi que ceux liés à la visualisation des données, 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.

Tout n'est toutefois pas *big* dans le *big data*. En effet, la grande majorité des collections de *big data* est caractérisée par le fait que certains types de données sont très présents alors que d'autres n'apparaissent que peu souvent ; on parle alors de « longue traîne », de « grand nombre d'événements rares » ou de « données déséquilibrées » pour qualifier ces types peu fréquents. Les lois de puissance sont en général retenues pour modéliser des distributions très inégale entre différents types, que ce soit pour de grandes collections textuelles (Yang *et al.*, 2003) ou pour des réseaux sociaux et plus généralement des réseaux complexes (*Barabási et al.,* 1999).

Enfin, au-delà de la distribution des traitements informatiques, un autre type de distribution voit le jour, celui sur les personnes au travers du phénomène de *crowdsourcing*. La plateforme *Amazon Mechanical Turk*[32] met ainsi en relation des personnes réalisant un certain nombre de tâches contre rémunération et d'autres désireuses de voir leurs tâches résolues. Le nombre de personnes à même de travailler sur la même tâche est potentiellement élevé et un grand nombre de problèmes de l'analyse de données liés au passage à l'échelle, comme l'annotation, le nettoyage, l'évaluation, trouve ainsi une réponse souvent raisonnable en termes de coûts et de délais. Toutefois, il est nécessaire pour réaliser ces opérations de résoudre un certain nombre de problèmes spécifiques qui ne sont pas toujours associés au big data mais peuvent être déterminants pour la mise en œuvre d'une solution à grande échelle. Ces problèmes concernent en particulier la détection d'experts au sein d'une communauté très large et la détermination de politiques d'échantillonnage à des fins d'annotation.

3.3. Extraction de connaissances

L'extraction de connaissances dans les données (ECD) est un processus défini comme un ensemble de tâches permettant la découverte automatique de motifs d'intérêts (ou connaissance) à partir de données. Ces motifs doivent être valides, potentiellement utiles et compréhensibles. Ce processus est couramment composé de trois étapes principales (Han *et al.*, 2006) :

- **Le pré-traitement.** Les données brutes sont souvent peu adaptées à la fouille car elles peuvent s'avérer bruitées et nécessitent l'application de quelques traitements pour les nettoyer, comme la suppression de valeurs aberrantes ou la prise en compte des données manquantes. Comme indiqué précédemment, le

[31] L'essor de ce domaine de recherche peut être en partie observé par l'augmentation du nombre de livres scientifiques dédiés.
[32] https://www.mturk.com/mturk/

volume permet souvent de s'affranchir de la nécessité de corriger des problèmes. Les données peuvent également être trop volumineuses horizontalement, i.e., comporter un nombre important de dimensions, et donc nécessiter l'application de techniques de réduction de dimensionnalité ou de détection de corrélations (Lai *et al.*, 2013)(Laporte *et al.*, 2014), ou verticalement, i.e., contenir de nombreux n-uplets, et donc nécessiter l'application de techniques d'échantillonnage (Cohen *et al .,* 2009).

- **La fouille de données.** Les algorithmes de fouille de données sont ensuite appliqués sur ces données nettoyées. Schématiquement, un algorithme de fouille de données a pour objectif d'exhiber un modèle pour décrire les données. La littérature regorge d'approches permettant la construction de ce modèle. Les décrire n'est pas l'objet de cet article. On peut néanmoins rapidement mentionner les deux grandes catégories d'approches existantes : les méthodes descriptives et les méthodes prédictives. Les méthodes descriptives permettent de simplifier et de mieux comprendre un jeu de données. Parmi les approches rentrant dans cette catégorie on peut citer les techniques de statistiques exploratoires multidimensionnelles (Lebart *et al.*, 1995), la classification non supervisée (Jain *et al.*, 1999) ou l'extraction de règles d'associations (Agrawal *et al.,* 1993). Les méthodes prédictives ont pour objectif d'apprendre un modèle d'une ou plusieurs variables cible à partir d'un jeu de données d'entraînements. Ces modèles sont ensuite utilisés pour prédire de nouvelles données dont la ou les variables cibles sont inconnues. Parmi les domaines de recherche rentrant dans cette catégorie, on peut citer la classification supervisée (Hengel et Rossiter, 2003), les techniques de régression (Var, 1998) ou encore la construction de modèle génératif (Bishop et Nasrabadi, 2006).

- **Le post-traitement.** La plupart du temps, les motifs extraits ne sont pas tous utiles. Il est alors nécessaire de filtrer ces motifs en fonction de l'application considérée et ensuite de proposer des techniques efficaces pour les visualiser correctement.

Ce processus est généralement itératif dans le sens où il faut souvent procéder à plusieurs raffinements successifs pour obtenir des résultats satisfaisants. Traditionnellement, les données utilisées pour la fouille sont structurées et stockées dans des bases de données relationnelles ou dans des fichiers textuels au format tabulaire (csv). Avec l'avènement du *big data*, et plus particulièrement le développement du *web* et l'explosion du volume de données textuelles disponibles, une sous-disciplines est apparue et ne cesse de gagner en popularité : le *web mining* (Liu, 2007). Schématiquement, le *web mining* regroupe des outils qui permettent d'analyser la structure du *web*, des réseaux sociaux ou de graphes de terrain, le contenu des pages *web* (souvent réduit au *text mining)* ou les habitudes de navigation des internautes.

Dans la suite de cette section, nous discutons de l'impact du *big data* sur ces trois étapes afin d'exhiber aussi bien l'évolution des traitements dans ce contexte difficile que les défis associés.

3.3.1. Le pré-traitement des données

Le pré-traitement des données est l'étape la plus consommatrice en temps dans un processus ECD. Elle est toutefois cruciale pour extraire une connaissance de qualité et doit donc nécessiter une attention particulière. Le *big data* impacte massivement ce pré-traitement de données. D'abord, dans le cas de données très volumineuses, la plupart des algorithmes de fouille de données peinent à passer à l'échelle. Des techniques de réduction de dimensionnalité ou d'échantillonnages sont alors tout à fait appropriées (Yan *et al.*, 2006). En outre, disposer de grandes quantités de données va souvent de pair avec une baisse de qualité de ces données, comme la présence accrue de valeurs aberrantes ou manquantes qu'il faut détecter et filtrer (Bollier *et al.*, 2010).

Ensuite, les algorithmes traditionnels de *data mining* peinent à intégrer des sources de données hétérogènes et réclament souvent en entrée un format prédéfini. Dans le contexte du *big data* et surtout lorsque les sources de données sont distribuées, l'intégration de données provenant de sources différentes se révèle difficile. Supposons par exemple que des données provenant d'un grand nombre de sites *web*, comme les agences de voyage en ligne, doivent être intégrées pour être ensuite analysées. Ces informations ne seront naturellement pas formatées de la même manière, rendant impossible leur simple concaténation. L'intégration de ces données requiert alors d'appliquer des techniques de réconciliation de schémas. Celles-ci opèrent habituellement sur des paires de schémas, ce qui est une limite à l'adoption de ces techniques dans notre contexte (He *et al.*, 2004). Dans le cas où un grand nombre de schémas doivent être réconciliés, il est possible d'utiliser des approches statistiques pour mettre en avant des corrélations (He *et al.*, 2004), des *clusters* (Wu *et al.*, 2004) entre schémas ou encore des interfaces permettant l'intégration de diverses sources web hétérogènes (Dragut *et al.*, 2009)(Li *et al.*, 2013).

Enfin, la vélocité des données impose que les pré-traitements effectués soient les plus rapides possibles. Lorsque les données sont si rapides qu'elles peuvent être modélisées sous la forme de flot, une contrainte supplémentaire s'ajoute : les données ne peuvent être lues qu'une seule fois, contrainte connue sous le terme *single pass constraint* (Gama et Gaber, 2007).

3.3.2. La fouille de données

Les algorithmes et les logiciels permettant la fouille de données doivent eux-mêmes être reconsidérés pour faire face aux particularités des *big data*. Lin et Ryaboy (2013) montrent que, compte tenu de l'état actuel des outils de fouille de données, il n'est pas aisé de réaliser des analyses sophistiquées lorsque les données sont très volumineuses. Ainsi, pour réaliser de telles analyses, il est nécessaire de consacrer un effort important sur l'adaptation des méthodes existantes pour les rendre robustes face à d'importants volumes de données. C'est dans ce contexte que sont apparues des déclinaisons *MapReduce* d'algorithmes de fouille de données, comme ceux permettant l'extraction de règles d'association ou le *clustering*. Pour mutualiser les efforts du monde académique dans cette direction, la librairie

Mahout[33] a été construite par la fondation *Apache* et met à disposition de nombreux algorithmes de fouille de données, d'apprentissage automatique ou de recherche d'information adaptés pour pouvoir traiter des données très volumineuses.

La variété des données est également une problématique sensible pour les outils de fouille de données. En effet, l'avènement du *web* a largement favorisé l'émergence d'un nouveau type de données : les données interconnectées, représentables sous la forme d'un graphe. Dès lors, l'analyse et la découverte de sous-structures particulières dans ces immenses graphes sont devenues une problématique majeure en fouille de données (Kang et Faloustos, 2013) depuis ces dernières années et toutes les plus grandes conférences du domaine possèdent aujourd'hui des sessions autour de la fouille de graphes. Plus intéressant encore, ces grands graphes sont la plupart du temps attribués, c'est-à-dire qu'il existe des propriétés attachées aux nœuds et/ou aux relations, et il fait également sens de croiser ces informations avec des données plus classiques et plus ou moins structurées, comme des données relationnelles ou des fichiers textes. Deux acteurs clés du domaine, Sun et Han (2013), ont récemment abordé ce point et ont discuté de la pertinence et des nombreux défis que représentent la fouille de graphe de données hétérogènes pour le développement du *big data mining* dans les années à venir.

Enfin, comme mentionné précédemment, le caractère véloce des données pousse à ne considérer chaque donnée apparaissant dans un flot qu'une seule fois. La fouille de flot de données commence à être une problématique de recherche mature et de nombreuses approches existent pour gérer de telles données (Gaber *et al.*, 2005). Cependant, lorsque l'on ajoute le volume, c'est à dire un nombre de flot très important, à la vélocité, les techniques existantes montrent leurs limites et de nouvelles techniques doivent alors être proposées (Bifet, 2013 ; Amatriain, 2013). Quelques logiciels libres existent pour mutualiser les efforts du monde académique dans cette direction, comme MOA (Bifet *et al.*, 2010) et SAMOA (De Francisci Morales, 2013).

3.3.3. Le post-traitement des données

La connaissance résultant d'un processus ECD doit pouvoir être compréhensible et interprétable par un expert du domaine des données. Ainsi, la mise au point de techniques de visualisation pour présenter le résultat d'un algorithme de fouille de données est particulièrement pertinente. La visualisation de résultats d'algorithmes de fouille de données a toujours été délicate car elle est fortement corrélée avec le type de connaissance extraite. Ainsi, s'il est clair que cette problématique est cruciale pour l'adoption de la fouille de *big data* dans le tissu industriel, très peu de recherches se focalisent dessus. Plus généralement, la visualisation de *big data* connaît néanmoins un essor important ces dernières années. Parmi les indicateurs qui permettent d'étayer cette affirmation on peut citer l'organisation du premier

[33] http://mahout.apache.org

atelier spécifiquement dédié à cette thématique en octobre 2013[34] ainsi que l'apparition de techniques de visualisation s'appuyant sur *Hadoop* (Vo *et al.*, 2011; Chen *et al.*, 2011).

4. Défis du *big data*

Comme le souligne le rapport du Cigref (2013), « la finalité du *big data* est d'améliorer l'efficacité des prises de décision et rendre l'ensemble de la chaîne de valeur plus efficace ». Cependant, donner du sens aux informations « dormantes » fait face à plusieurs défis : créer une démarche d'entreprise et intégrer des compétences plurielles, considérer la qualité des données et les aspects éthiques.

4.1. Démarche d'entreprise

Le *big data* touche a priori tous les secteurs puisque tous sont soit producteur, soit collecteur de données, soit les deux. Le taux de pénétration des données numériques dans les différents secteurs présenté dans l'étude de McKinsey l'illustre bien (voir Figure 2).

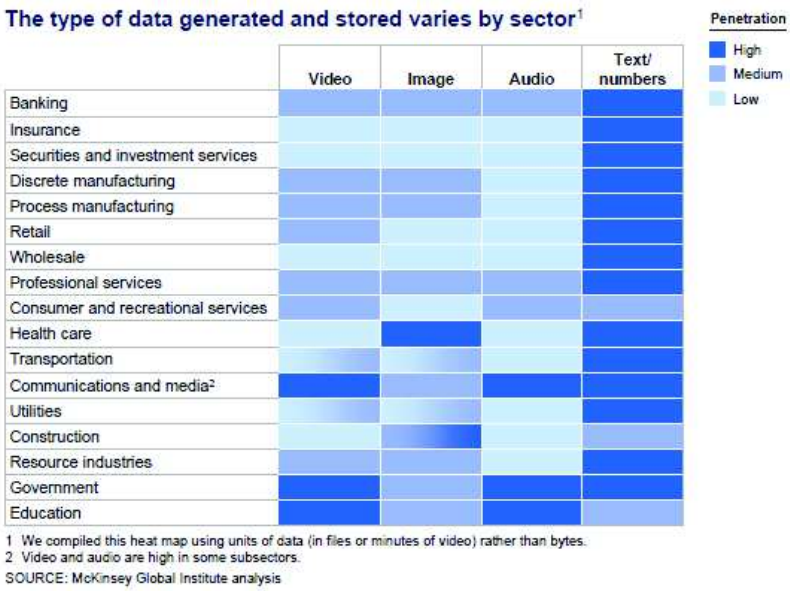

Figure 2 : Type de données générées et stockées par secteurs (extrait de Mc Kinsey)

Dans la mesure où les données sont produites ou collectées par tous les services d'une organisation, un enjeu majeur est la prise de conscience de la valeur potentielle des données manipulées, non pas nécessairement prises de façon isolées mais plutôt combinées avec d'autres données internes ou externes à l'organisation.

[34] http://vis.ucdavis.edu/Workshops/BigDataVis2013/

Une démarche *big data* implique donc un décloisonnement des métiers. Cette démarche peut avoir deux grands types d'objectifs : améliorer la qualité du produit ou du service et créer de nouvelles opportunités. L'amélioration de la qualité fait référence en particulier à une meilleure réponse au client par l'analyse de leur comportement ou de leur goût ou à une réponse plus rapide aux pannes, dysfonctionnements ou signaux faibles. Les nouvelles opportunités résultent plus d'une meilleure prise en compte de l'environnement politique, sociétal ou culturel. Il s'agit là également d'une des motivations de l'ouverture des données (*open data*) pour lesquelles l'innovation permet de créer de la valeur. Plusieurs types d'applications basées sur les données de géo-localisation ont d'ailleurs vu le jour.

Dans cette démarche, l'intégration des *big data* dans le système d'information n'est pas triviale. Une des étapes cruciale est de prendre conscience de la valeur des données internes. Le système d'information peut jouer un rôle dans cette prise de conscience mais il n'en est pas le moteur. « Le rôle de la direction générale et des métiers est clé sur ce sujet, car ce sont eux les seuls à même de définir les informations utiles à l'amélioration de leurs prises de décisions » (Cigref). Parallèlement, il s'agit d'identifier les données externes potentiellement utiles (leur valeur potentielle), de les collecter et de les intégrer (en prenant en compte les volume, vélocité, variété et véracité).

Du point de vue des ressources, outre les ressources matérielles et logicielles, le *big data* nécessite également de nombreuses compétences : informatique pour la gestion et le traitement des données, mathématique pour leur analyse adaptée, juridique, économique, … en plus de l'expertise métier. Toutes ces compétences n'existent pas forcément en interne dans l'organisation. Ainsi, la démarche *big data* peut nécessiter de la formation ou de l'expertise externe pour mener à bien le projet.

De nombreux projets pilotes voient le jour dans de nombreux secteurs variés ; le guide du *big data* (2013) fournit ainsi des exemples variés : indicateurs en lien avec les plis à *La Poste* ou personnalisation des courriers chez *Monoprix* et *dunnhumby*. *SQLI*, entreprise spécialisée dans l'accompagnement des entreprises dans la mise en œuvre de plateformes digitales orientées performance de l'entreprise, préconise un cycle composé des étapes suivantes : (1) déterminer ce que l'on cherche, (2) appréhender les sources de données pouvant apporter des réponses, (3) maîtriser l'origine et analyser ces sources, (4) croiser les résultats des différentes sources et (5) consolider et interpréter les résultats. Cisco, quant à lui, indique qu'à l'heure actuelle les projets *big data* restent ciblés sur des échantillons de données pour la majeure partie des entreprises[35]. D'ailleurs, dans son livre, Ohlhorst (2012) préconise de « commencer petit » ! Beaucoup s'accordent à penser que le *big data* est très prometteur ; le *big data* a d'ailleurs été identifié par l'ancienne présidente d'Areva, Anne Lauvergeon, comme un des sept secteurs stratégiques d'innovations technologiques et industrielles pour la France à l'horizon 2030. Pour l'heure, les entreprises *Business to Consumer* semblent plus engagées dans cette démarche. Le

[35] http://gblogs.cisco.com/fr-datacenter/2013/10/14/big-data-la-vision-des-grandes-entreprises-francaise-par-le-cigref/

frein majeur concerne l'investissement nécessaire et une méconnaissance du ROI (*Return On Investment*).

Le ROI, que certains déclinent en *Return On Information*, est difficile à estimer malgré le grand succès des entreprises pionnières en la matière. *Google* qui a débuté en janvier 1996 possédait 50 milliards de dollars de chiffre d'affaire en 2012. Les modèles économiques des grosses entreprises d'Internet sont basés sur les revenus publicitaires, difficilement transposables dans tous les secteurs.

4.2 Qualité des données et préservation de la vie privée

4.2.1. Qualité des données

Disposer de données de qualité a toujours été une préoccupation importante dans un contexte décisionnel (Redman et Blanton, 1997). En effet, une décision ne peut être valide que si elle a été prise à partir de données qui représentent aussi fidèlement que possible le monde réel qui bénéficiera de cette décision. A ce titre, on peut légitimement se demander si plus de données correspond nécessairement à des données de meilleures qualités (Boyd et Crawford, 2012). Par exemple, de nombreux travaux de recherche autour du *big data* utilisent les données extraites du réseau social *Twitter*[36]. Il convient alors d'être extrêmement vigilant et de ne pas faire l'amalgame entre utilisateurs de *Twitter* (la population représentée par le jeu de données) et population « réelle ». Plusieurs arguments permettent d'étayer cette précaution à prendre. D'abord, l'API qui permet de récupérer le flot de données de *Twitter* ne permet de ne récupérer qu'un échantillon des données émises, sans que la constitution de cet échantillon soit clairement détaillée. Ensuite, même si ce réseau social tend à devenir de plus en plus populaire, il ne touche qu'une certaine frange de la population.

Ainsi, aussi volumineuse que soient ces données, il convient de ne pas les considérer comme des données de qualité si l'on souhaite, par exemple, les utiliser comme support pour construire une campagne publicitaire nationale à destination de certaines couches de la population dont la représentation sur *Twitter* n'est pas avérée.

Cet exemple illustre parfaitement un fait depuis longtemps établi : il n'existe pas de donnée universellement de qualité. En effet, si l'on peut donner une liste de critères pour décrire ce qu'est une donnée de qualité, comme la cohérence, la précision, la confiance attribuée, la complétude ou encore la fraîcheur, il est important de noter que le processus de qualification des données est fortement dépendant du but poursuivi par l'utilisation de ces données (Strong *et al.*, 1997). Les données mentionnées précédemment auraient ainsi pu être considérées comme de grande qualité si la cible visée par la campagne publicitaire avait été « les utilisateurs de *Twitter* ». Une différence fondamentale différencie alors l'analyse de données « classiques » de l'analyse de *big data*. Généralement, la connaissance, ou au moins le type de connaissance, que l'on souhaite extraire est déterminé à

[36] https://twitter.com

l'avance. Le prétraitement des données ainsi que leur intégration dans le système d'information, par le biais de processus ETL, est alors prédéterminé par l'objectif à atteindre (Vassiliadis *et al.*, 2002). Dans le contexte *big data*, cette connaissance *a priori* des méthodes qui seront appliquées est beaucoup plus floue. La stratégie est alors de collecter autant de données que possible, sans prétraitement, et, ensuite, de sélectionner un sous-ensemble de ces données à des fins d'analyse. Nous détaillons ci-dessous quelques arguments abondant dans la direction de la contextualisation du processus de qualification de données dans le cadre des *big data*:

- Considérer le type de données manipulées : la variété des données induit nécessairement de s'adapter aux différents types de données (Rahm et Do, 2010). Ainsi, si les données considérées sont de type transactionnel, comme des données d'achats ou de clients, l'application de techniques existantes adaptées à ce type de données peut être envisagée pour peu que ces techniques puissent affronter efficacement l'aspect volume des données. Par contre, les données produites par des machines, comme des logs ou des capteurs, réclament des traitements particuliers. En effet, ces données sont souvent considérées de faible qualité car peu précises ou contenant fréquemment des valeurs manquantes ou aberrantes. Un autre cas particulier concerne les données issues des réseaux sociaux. Ces données présentent une double particularité. Dans un premier temps, l'analyse des données échangées requiert bien souvent l'utilisation et l'adaptation de techniques provenant de l'analyse de données textuelles. Ensuite, les métadonnées associées sont cruciales dans l'intégration de données sociales à des données plus classiques, comme par exemple des fiches clients. Déterminer la confiance que l'on attribue à ces métadonnées et évaluer leur validité apparaissent alors comme des tâches de qualification des données cruciales pour pouvoir intégrer ces données sociales et assurer la cohérence de l'ensemble du système d'information ;

- Toutes les analyses ne requièrent pas le même niveau de qualité dans les données : ainsi, si l'analyse consiste à extraire des motifs représentant une connaissance générale d'un ensemble de données, la présence de données aberrantes n'impactera que marginalement le processus (Jansen *et al.*, 2009). Par exemple, l'analyse des *logs* d'un site marchand pour extraire les chemins de navigation ayant le plus conduit à un acte d'achat n'est que très peu concernée par le nettoyage des données.

Enfin, les données de mauvaise qualité peuvent se révéler intéressantes dans l'analyse (Li *et al.*, 2010). Ce qui peut apparaître comme du bruit ou une transaction aberrante peut se révéler être de la plus haute importance. Par exemple, de telles données peuvent être révélatrices d'une fraude ou d'une défaillance dans le système de mesures.

4.2.2. Préservation de la vie privée

Schématiquement, on peut dire que le *big data* traite de « comment collecter et analyser le plus de données possible » (et généralement sur les utilisateurs et leurs habitudes). D'un autre côté, la préservation de la vie privée s'apparente plus à

« comment ne pas diffuser des données concernant les utilisateurs ». Ces deux enjeux semblent alors en contradiction et il est donc naturel que des difficultés se posent quant à la préservation de la vie privée dans le contexte du *big data*. Quelques exemples de corruption de la vie privée font référence pour justifier la mise au point de techniques efficaces et d'un cadre juridique pour se protéger des dérives associées au *big data*. Nous en mentionnons deux notables.

En 2006, un groupe de recherche d'Harvard a collecté les profils *Facebook* de 1700 étudiants afin d'étudier comment leurs intérêts et interactions évoluent avec letemps (Lewis et al. 2008). Ces données, anonymisées au préalable, ont été diffusées librement pour permettre aux autres chercheurs de les utiliser. D'autres équipes ont rapidement montré qu'il était possible de désanonymiser une partie du jeu de données (Zimmer, 2008).

Le second exemple provient de *Netflix*, un service de location de contenu vidéo en ligne très populaire. *Netflix* dispose d'un service de recommandation, nommé *Cinematic*, qui propose des contenus aux utilisateurs en fonction des préférences formulées par l'utilisateur. *Netfix* a mis à disposition ses données en octobre 2006 à l'occasion d'un challenge appelé « *Netflix Prize* ». Son objectif était d'améliorer l'efficacité du système de recommandation. Narayanan et Shmatikov (2006) ont montré que 96% des utilisateurs peuvent être identifiés de manière unique en ne considérant seulement qu'au plus 8 films évalués par l'utilisateur avec la date d'évaluation.

La protection de la vie privée devient de plus en plus importante quand les données sont massives, arrivent à un débit souvent très élevé et sont stockées sur de nombreux sites distants. Assez logiquement, l'augmentation du volume de données multiplie les risques de constater des fuites de données sensibles. De plus, le stockage des données dans les nuages augmente également les risques qu'une personne malicieuse obtienne les données sur un individu ou une organisation. En effet, dans la mesure où les données sont stockées et répliquées sur de nombreux sites distants, la probabilité qu'un attaquant exploite une faille de sécurité d'un des sites où sont hébergées les données augmente. La protection de la vie privée devient donc un sujet encore plus sensible dans le contexte du *big data*. Mais alors, comment s'en prémunir ? Les deux grandes directions considérées de nos jours pour protéger les utilisateurs de ces dérives sont la mise en place de garde-fous juridiques et le développement de technologies spécifiques. Nous détaillons brièvement les défis associés à ces deux directions.

La tendance actuelle est de stocker la plupart des *big data* dans les nuages (Manyika *et al.*, 2011). Un intérêt croissant de la communauté se porte donc sur comment concilier préservation de la vie privée et stockage et traitement des données dans les nuages. Les défis techniques sont nombreux. D'abord, les données stockées ne sont en théorie consultables que par le propriétaire des données ou des utilisateurs qu'il aura, au préalable, autorisé à accéder aux données. Une première série de défis traitent de comment s'assurer que l'accès aux ressources ne soit garanti que pour les utilisateurs qui en ont effectivement le droit (Yu *et al.*, 2010). Ensuite, la virtualisation est de plus en plus utilisée pour héberger des services. Si

cette stratégie est intéressante d'un point de vue de l'administration, elle offre néanmoins une plus grande surface d'attaque. Des mécanismes pour assurer l'isolation des machines virtuelles et sécuriser leurs communications sont alors nécessaires (Ormandy, 2007). Enfin, un utilisateur ou une organisation peut recourir à plusieurs services offerts par différents fournisseurs. Par exemple, les données peuvent être stockées par X, puis un premier service hébergé par Y est appliqué et le résultat est ensuite envoyé au service hébergé par Z pour être visualisé. Ainsi, de la multiplication des échanges et de la différence des systèmes de protection naît un risque accru de faille de sécurité et donc de divulgation de la vie privée. La mise au point de modèles pour pallier ces possibles axes d'attaque représente alors un enjeu important (Takabi *et al.*, 2010).

Enfin, la protection de la vie privée ne saurait trouver ses solutions dans des considérations techniques seulement. C'est également un problème sociétal dont les institutions s'emparent petit à petit. Cependant, nous sommes forcés de constater qu'il existe encore des disparités fortes entre les différents pays. Par exemple, en Europe, le droit à la protection de la vie privée est directement inscrit dans la convention européenne des droits de l'Homme (article 8) pour prémunir les organisations publiques de l'utilisation abusive de données privées. Cependant, dans la mesure où les organisations publiques ne sont pas les seules à potentiellement pouvoir abuser de la vie privée des individus, la commission européenne a mis au point la « convention pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel[37] » en 1981. A l'inverse, les Etats-Unis sont beaucoup plus permissifs concernant cet aspect. Il n'existe pas de loi qui encadre complètement la collecte, le stockage et l'utilisation de données privées et quiconque saisit des données, même si elles sont retenues sans autorisation, a le droit de les stocker et de les utiliser. Néanmoins, pour faciliter les échanges commerciaux avec les pays européens, le ministère américain de l'économie a développé le « *Safe Harbor Arrangment* ». Face à cette hétérogénéité des règlements à travers le monde, les institutions internationales doivent accroître leurs efforts pour uniformiser le droit à la protection de la vie privée.

5. Conclusion

De nombreux défis sont associés au *big data* découlant du volume, de la variété, de la vélocité, de la véracité et de la valeur des données (Kuntze , 2008) (Laoutaris , 2010). Nous en avons abordé un certain nombre ici que nous résumons ci-dessous.

Le **stockage des données** fait écho au problème de sécurité. La sécurité a trait à différents aspects qui garantissent la sécurité des données et de l'équipement (des salles et des bâtiments adaptés aux processus de réplication ou duplication). D'après l'ISO27001, la sécurité se réfère à la confidentialité, la disponibilité, l'intégrité, la contrôlabilité. Par exemple, la confidentialité doit être considérée différemment en fonction des données : certaines peuvent être publiques, d'autres sont privées ; ce

[37] http://conventions.coe.int/Treaty/FR/Treaties/Html/108.htm

qui soulève également des questions juridiques. L'évolution élastique des ressources est une question spécifique liée au *cloud*, qui permet aux utilisateurs de ne payer que ce qu'ils utilisent. Cette question est également liée à la question du *Green IT* dans la mesure où les équipements sont escaladés sur demande, ne consomment que lorsqu'ils sont allumés et sont partagés (Armbrust, 2009).

La **description des données** est un élément essentiel en particulier dans le cadre de données partagées, qu'elles soient ouvertes ou non. L'utilisation et la ré-utilisation correcte des données impliquent qu'elles soient bien décrites (méthode de production ou de collecte, description des contenus, pré-traitements éventuels). Les métadonnées sont un moyen de décrire les données de façon compréhensible pour différentes communautés d'utilisateurs. Certaines communautés sont assez avancées dans la description de données, mais d'autres devront élaborer des processus et des cadres pour décrire leurs données pour une utilisation à long terme. Par exemple, en astronomie, le projet *Observatoire Virtuel* (http://www.ivoa.net/) vise déjà depuis quelques années à déployer des normes internationales pour la description des données et pour définir des protocoles d'accès appropriés à des fins d'interopérabilité. Néanmoins, l'accessibilité de ces métadonnées pour les nouvelles communautés scientifiques soulève des questions non résolues, comme la réutilisation et l'interprétation.

Le **pré-traitement des données** : un autre défi est lié au fait que certaines données brutes sont rarement utilisées sans pré-traitements (par exemple, les données observées des instruments). Des outils de visualisation qui fournissent des vues globales des données peuvent également être d'une grande aide pour leurs utilisateurs (Ware, 2004). Les différents niveaux de traitement des données et leur visualisation doivent être définis pour de nouveaux projets, en tenant compte des besoins des utilisateurs, des capacités de traitement actuelles et de celles qui seront disponibles dans un proche avenir.

Les **exigences techniques** : beaucoup de projets portant sur de grandes quantités de données (entrepôts de données) sont des échecs relatifs, car ils ne répondent pas aux besoins de leurs utilisateurs (Giorgini *et al.*, 2008). Les exigences des méthodes d'ingénierie utilisées pour les systèmes d'information classiques doivent être adaptées de manière significative pour la conception de systèmes basés sur des *big data*.

L'**extraction de connaissance**: le traitement de l'information afin de découvrir les modèles et les règles entre les données ou pour aider les décisions sont un objectif essentiel dans de nombreuses branches de l'industrie, de la recherche ainsi que dans les sciences appliquées. Cela implique de considérer des méthodes de fouille de données, de compression de données, de synthèse d'information. Diverses données peuvent être extraites et calculées à partir des mêmes données brutes, sous divers formats, en fonction des objectifs ou de la science pour laquelle elles sont utilisées. Enfin, la connaissance doit être extraite de ces données pour que les organisations soient en mesure de les exploiter à travers différentes modalités et langues. Le défi général est d'aider les scientifiques, les chercheurs, les industriels et

les décideurs en apportant des solutions innovantes en matière d'aide ou de « co-travail » pour produire des résultats.

Enfin, la **réduction de la durée de temps entre la capture et le traitement** est cruciale. Même si les données sont bien décrites, leur volume et leur complexité peuvent conduire à des traitements longs. Dans l'intervalle, des volumes de nouvelles données sont collectés ou produits. Par ailleurs, tous les traitements ne sont pas pensés au moment de la collecte des données. En conséquence, le traitement des données peut être réalisé 10 ou 20 ans après leur capture. L'un des défis ici est de réduire ce délai en guidant les utilisateurs dans l'utilisation des données et en les orientant vers les données particulièrement significatives par exemple.

Bibliographie non numérotée et références

Agrawal R., Imieliński T. et Swami A. (1993). Mining association rules between sets of items in large databases. In *ACM SIGMOD Record* (Vol. 22, No. 2, pp. 207-216). ACM.

Agrawal D., Das S. et El Abbadi A. (2011). 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.

Allan J. to Zhang Y. (44 auteurs) (2012) Frontiers, Challenges, and Opportunities for Information Retrieval, Second strategic workshop on information retrieval in Lorne.

Alpert J. et Hajaj N. (2008) Software Engineers, Web Search Infrastructure Team , Google official blog, Posted: Friday, July 25, 2008 http://googleblog.blogspot.fr/2008/07/we-knew-web-was-big.html.

Amatriain X. (2013). Mining large streams of user data for personalized recommendations. *ACM SIGKDD Explorations Newsletter*, *14*(2), 37-48.

Amini M.-R et Gaussier E. (2013). *Recherche d'information*, Applications, modèles et algorithmes, Eyrolles, ISBN 978-2-212-13532-9.

Arens Y., Chee C. Y., Hsu C. N. et Knoblock C. A. (1993). Retrieving and integrating data from multiple information sources, *International Journal of Intelligent and Cooperative Information Systems*, volume 2, n° 02, p. 127-158.

Armbrust M., Fox A., Griffith R., Joseph A. D., Katz R., Konwinski A., ... et Zaharia M. (2010). A view of cloud computing. *Communications of the ACM*, 53(4), 50-58.

Barabási A.L. et Albert R. (1999). Emergence of Scaling in Random Networks. Science, Vol. 286, No. 5439. pp. 509-512.

Barbieri, N., Bonchi, F., et Manco, G. (2013, February). Cascade-based community detection. *In Proceedings of the sixth ACM international conference on Web search and data mining* (pp. 33-42). ACM.

Berners-Lee T. (2006). Linked Data - Design Issues. Retrieved July 23, http://www.w3.org/DesignIssues/LinkedData.html

Bifet A. (2013). Mining Big Data in Real Time. *Informatica*, *37*, 15-20.

Bifet A., Holmes G., Kirkby R. et Pfahringer B. (2010). Moa: Massive online analysis. *The Journal of Machine Learning Research*, *99*, 1601-1604.

Bishop C. M. et Nasrabadi N. M. (2006). *Pattern recognition and machine learning* (Vol. 1, p. 740). New York: springer.

Bizer C. et Cyganiak R. (2006, November). D2r server-publishing relational databases on the semantic web. *In 5th international Semantic Web conference* (p. 26).

Bollier D. et Firestone C. M. (2010). *The promise and peril of big data*. Washington, DC, USA: Aspen Institute, Communications and Society Program.

Boyd D. et Crawford K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662-679.

Brewer, E. (2010). A certain freedom: thoughts on the cap theorem. *In Proceedings of the 29th ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (pp. 335-335). ACM.

Brutlag J. D., Hutchinson H. et Stone M. (2008). User Preference and Search Engine Latency, *Quality and Productivity Research Conference*.

Carroll R., (2013), Welcome to Utah, the NSA's desert home for eavesdropping on America, *The guardian*, 14 juin.

Cattell R. (2011). Scalable SQL and NoSQL data stores. *ACM SIGMOD Record*, 39(4), 12-27.

Chen K., Xu H., Tian F. et Guo, S. (2011). Cloudvista: Visual cluster exploration for extreme scale data in the cloud. In *Scientific and Statistical Database Management* (pp. 332-350). Springer Berlin Heidelberg.

Chen H., Chiang R.H.L. et Storey V.C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, volume 36, n° 4, p. 1165–1188.

Cigref (2013). Big Data: la vision des grandes entreprises – *Opportunités et enjeux*, octobre.

Clarke C., Craswell N. et Soboroff I. (2005). Overview of the TREC 2004 Terabyte Track, NIST Special Publication: SP 500-261.

Clarke C., Craswell N., Soboroff I. (2010). Overview of the TREC 2009 Web Track, NIST Special Publication: SP 500-278.

Cleverdon C. W., Mills J. et Keen E. M. (1966). Factors determining the performance of indexing systems. Cranfield, UK: Aslib Cranfield Research Project, College of Aeronautics. (Volume 1:Design; Volume 2: Results).

Cohen J., Dolan B., Dunlap M., Hellerstein J. M. et Welton C. (2009). MAD skills: new analysis practices for big data. *Proceedings of the VLDB Endowment*, *2*(2), 1481-1492.

Cox M. et Ellsworth D. (1997). Managing Big Data for Scientific Visualization, In ACM SIGGRAPH, Course 4, Exploring Gigabyte Datasets in Real-Time: Algorithms, Data Management and Time-Critical Design. http://www.dcs.ed.ac.uk/teaching/cs4/www/visualisation/lectures/98-99/lect16ref.ps.gz

Cormode G. et Muthukrishnan S. (2005). Effective computation of biased quantiles over data streams. *In Proceedings of the 21st International Conference on Data Engineering, ICDE 2005*, 5-8 April 2005, Tokyo, Japan.

Davenport T. H., Barth, P. et Bean R. (2012). How 'Big Data'is Different. *MIT Sloan Management Review*, volume 54, n°1, p. 22-24.

De Francisci Morales G. (2013). SAMOA: a platform for mining big data streams. In *Proceedings of the 22nd international conference on World Wide Web companion* (pp. 777-778). International World Wide Web Conferences Steering Committee.

Dkaki T., Dousset D. et Mothe J. (1997). Mining Information in order to Extract Hidden and Strategical Information, *5th International Conference on Computer Assisted Information Retrieval, RIAO*, p. 32-51.

Dragut E. C., Kabisch T., Yu C., et Leser U. (2009). A hierarchical approach to model web query interfaces for web source integration. *Proceedings of the VLDB Endowment*, *2*(1), 325-336.

Editorial (2008). Community cleverness required, Nature, Special issue on Big Data, Volume 455, Issue 7209, p. 1.

El Haddadi A., Dousset B. et Berrada I. (2012). Establishment and application of Competitive Intelligence System in Mobile Device. *Journal of Intelligence Studies in Business*, volume 1, n°1, p. 87-96.

El Haddadi A., Dousset B. et Berrada I. (2011). Discovering patterns in order to detect weak signals and define new strategies. *Pattern Discovery Using Sequence Data Mining: Applications and Studies*. Kumar Pradeep (Eds.), IGI Global, p. 195-211.

e-skills UK (2013). Big Data Analytics: Adoption and Employment Trends, 2012–2017.

Farrance R. (2006). Timeline: 50 Years of Hard Drives, *PCWorld*, 13 Sept.

Gaber M. M., Zaslavsky A. et Krishnaswamy S. (2005). Mining data streams: a review. *SIGMOD Rec.* 34, 2 (June 2005), 18-26. DOI=10.1145/1083784.1083789 http://doi.acm.org/10.1145/1083784.1083789.

Gama, J., & Gaber, M. M. (Eds.). (2007). *Learning from data streams: processing techniques in sensor networks*. Springer.

Gigya (2013). 5 Ways Marketing Will Change in the Next 5 Years, *www.gigya.com/5-ways-big-data*.

Google (2013). Our history in depth, http://www.google.com/about/company/history/.

Grefenstette G. (1998). Cross-Language Information Retrieval, *The Information Retrieval Series*, Vol. 2, Kluwer Academic Publishers.

Han J., Kamber M. et Pei J. (2006). *Data mining: concepts and techniques*. Morgan kaufmann.

Harman D. (1993) Overview of TREC-1, *NIST Special Publication* 500-207, p. 1-20.

He B., Chang K. C. C. et Han J. (2004). Discovering complex matchings across web query interfaces: a correlation mining approach. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 148-157). ACM.

Hengl T. et Rossiter,D. G. (2003). Supervised landform classification to enhance and replace photo-interpretation in semi-detailed soil survey. *Soil Science Society of America Journal*, *67*(6), 1810-1822.

Holmstrom J. E. (1948). Section III. Opening Plenary Session, *The Royal Society Scientific Information Conference*, 21 Juin-2 Juillet 1948, London: Royal Society.

Jain A. K., Murty M. N. et Flynn P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, *31*(3), 264-323.

Jansen B. J., Spink A. et Taksai I. (2009). *Handbook of research on web log analysis*. London: Information Science Reference.

Kang U. et Faloutsos C. (2013). Big graph mining: algorithms and discoveries. *ACM SIGKDD Explorations Newsletter*, *14*(2), 29-36.

Kass G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied statistics*, 119-127.

Kim J., Hwang M., Jeong D.H. et Jung H. (2012). Technology trends analysis and forecasting application based on decision tree and statistical feature analysis, *Expert Systems with Applications, Elsevier*, volume 39, n°6, p. 12618–12625.

Kontostathis A., Galitsky L. M., Pottenger W. M., Roy S. et Phelps D. J. (2003). A survey of Emerging Trend Detection in Textual Data Mining, *Survey of Text Mining*, p. 185-224.

Kossmann D., Kraska T. et Loesing S. (2010). An evaluation of alternative architectures for transaction processing in the cloud. In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data (pp. 579-590).

Laffargue B. (2013). *Corp Events, Guide du Big Data, L'annuaire de référence à destination des utilisateurs*

Lai H.-J., Pan Y., Tang Y., and Yu R., Fsmrank: Feature selection algorithm for learning to rank, IEEE Trans. Neural Netw. Learning Syst., volume 24, n° 6, p. 940–952, 2013.

Laney D. (2001). 3d Data managment: controlling data volume, velocity and variety, *Appl. Delivery Strategies Meta Group* (949).

Laporte L., Flamary R., Canu S., Dejean S. et Mothe J., Nonconvex Regularizations for Feature Selection in Ranking With Sparse SVM, *Neural Networks and Learning Systems, IEEE Transactions on* , vol.PP, no.99,

La Recherche (2013). Big Data, N°482, Décembre.

Leavitt N. (2010). Will NoSQL databases live up to their promise? *Computer*, 43(2), 12-14.

Lebart L., Morineau A. et Piron M. (1995). *Statistique exploratoire multidimensionnelle* (Vol. 3). Paris: Dunod.

Lewis K., Kaufman J., Gonzalez M., Wimmer A. et Christakis N. (2008). Tastes, ties, and time: A new social network dataset using Facebook. com. *Social networks*, *30*(4), 330-342.

Li X., Li Z., Han J. et Lee J. G. (2009). Temporal outlier detection in vehicle traffic data. In *Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on* (pp. 1319-1322). IEEE.

Li Y., Wang Y., Jiang P., et Zhang Z. (2013). Multi-objective optimization integration of query interfaces for the Deep Web based on attribute constraints. *Data & Knowledge Engineering*, *86*, 38-60.

Lin J. et Ryaboy D. (2013). Scaling big data mining infrastructure: the twitter experience. *ACM SIGKDD Explorations Newsletter*, *14*(2), 6-19.

Liu B. (2007). *Web data mining: exploring hyperlinks, contents, and usage data*. Springer.

McAfee et Brynjolfsson (2012). Big Data : the Management Revolution, *Harvard Business Review.*

McWilliam G. (2000). Building stronger brands through online communities. *MIT Sloan management review*, volume 41, n°3.

Macdonald C., Tonellotto N. et Ounis I. (2012). Learning to predict response times for online query scheduling. *International ACM SIGIR conference on Research and development in information retrieval*, p. 621-630.

Manyika J., Chui M., Brown B., Bughin J., Dobbs R., Roxburgh C., Hung Byers A. (2011). Big data, the next frontier for innovation, competition, and productivity, McKinsey Global Institute.

Minelli M., Chambers M., Dhiraj A. (2012). *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses*. Wiley.

Mothe J. et Sahut G. (2011). Is a relevant piece of information a valid one? Teaching critical evaluation of online information. Approaches to Teaching and Learning Information Retrieval. Efthimiadis E. N., Fernández Luna J. M., Huete J., MacFarlane A. (Eds.), Springer, volume. 31, p. 153-168.

Mothe J. (2000). *Recherche et exploration d'informations -Découverte de connaissances pour l'accès; à l'information.* Habilitation à diriger des recherches, Université Paul Sabatier.

Narayanan A. et Shmatikov V. (2006). How to break anonymity of the netflix prize dataset. *arXiv preprint cs/0610105*.

Niranjan Mysore R., Pamboris A., Farrington N., Huang N., Miri P., Radhakrishnan, S., & Vahdat, A. (2009, August). PortLand: a scalable fault-tolerant layer 2 data center network fabric. In ACM SIGCOMM Computer Communication Review (Vol. 39, No. 4, pp. 39-50).

Ohlhorst F. J. (2012). *Big Data Analytics: Turning Big Data Into Big Money*. John Wiley & Sons.

Ormandy T. (2007). An empirical study into the security exposure to hosts of hostile virtualized environments. In *Proceedings of CanSecWest Applied Security Conference*.

Pentland A. (2013). Orienter la société grâce aux données massives, La Recherche, volume novembre 2013, n°433, p. 56-61.

Perrière G. (2013). Génomes, protéomes et transcriptomes à foison, La recherche.

Peters C., Braschler M. et Clough P. (2012). *Multilingual Information Retrieval*, Springer, 217 p. ISBN 978-3-642-23008-0.

Philip Chen C.L. et Zhang C.Y (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, *Information Sciences*, Available online 21 January 2014, ISSN 0020-0255, http://dx.doi.org/10.1016/j.ins.2014.01.015.

Pour la Science (2013). Big Bang numérique, N° Spécial, Nov.Pritchett D. (2008). Base: An acid alternative. Queue, 6(3), 48-55.

Pritchett D. (2008). Base: An acid alternative. Queue, 6(3), 48-55.

Rahm E. et Do H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, *23*(4), 3-13.

Redman T. C. et Blanton A. (1997). *Data quality for the information age*. Artech House, Inc..

Resnick P. et Varian H. R. (1997). Recommender systems. *Communications of the ACM*, 40(3), 56-58.

Ricci F., Rokach L., Shapira B. et Kantor P.B., (2011). *Recommender Systems Hadbook*.

Russom P. (2011). Big data analytics, TDWI Best Practices Report, Fourth Quarter.

Soutou C. (1994). Contribution à la conception d'une base de données fédérée : dérivation, évolution et intégration de schémas, Thèse de doctorat, Université P. Sabatier, Toulouse.

Sadre, R. et Haverkort, B.: (2008). Changes in the Web from 2000 to 2007. Managing Large-Scale Service Deployment, *Proceedings of the 19th IFIP/IEEE International Workshop on Distributed Systems (DSOM 2008)*, Volume 5273 of LNCS., Springer, p. 136–148.

Salton G., (1965). Progress in automatic information retrieval, *IEEE Spectrum*. Août, p. 90-103.

Sanderson M. et Croft W. B. (2012). The History of Information Retrieval Research. *Proceedings of the IEEE*, vol. 100, n° 13, p. 1444-1451.

Sauermann L., Cyganiak R. et Völkel M. (2011). Cool URIs for the semantic web.

Savoy J. (2005). Comparative study of monolingual and multilingual search models for use with asian languages. *ACM Transaction Asian Langage Information Process.* vol. 4, n°2, p. 163-189.

Shute J., Vingralek R., Samwel B., Handy B., Whipkey C., Rollins E., ... et Apte H. (2013). F1: A distributed SQL database that scales. *Proceedings of the VLDB Endowment*, 6(11), 1068-1079.

Strong D. M., Lee Y. W. et Wang R. Y. (1997). Data quality in context. *Communications of the ACM*, *40*(5), 103-110.

Sun Y. et Han, J. (2013). Mining heterogeneous information networks: a structural analysis approach. *ACM SIGKDD Explorations Newsletter*, *14*(2), 20-28.

Takabi H., Joshi J. B. et Ahn G. J. (2010). Security and privacy challenges in cloud computing environments. *Security & Privacy, IEEE*, *8*(6), 24-31.

The Thomson Corporation, (2007). Global Patent Sources: An Overview of International Patents, ISBN: 978 1 905935 07 9.

Untersinger M. (2013). Prism, un accès privilégié aux serveurs des géants de l'Internet, *Le Monde*, 22 oct.

Upson J. W., Ketchen D. J., Connelly B. L., Ranft A. L. (2012). Competitor analysis and foothold moves. *Academy of Management Journal*, volume 55, n°1, p. 93-110.

Var I. D. (1998). Multivariate data analysis. *vectors*, *8*, 6.

Vassiliadis P., Simitsis A. et Skiadopoulos S. (2002). Conceptual modeling for ETL processes. In *Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP* (pp. 14-21). ACM.

Velte T., Velte A. et Elsenpeter R. (2009). Cloud computing, a practical approach. McGraw-Hill, Inc..

Vo, H. T., Bronson, J., Summa, B., Comba, J. L., Freire, J., Howe, B., ... & Silva, C. T. (2011, October). Parallel visualization on large clusters using MapReduce. In *Large Data Analysis and Visualization (LDAV), 2011 IEEE Symposium on* (pp. 81-88). IEEE.

VoltDB L. L. C. (2010) VoltDB Technical Overview, Whitepaper.

W3C (2000) A Little History of the World Wide Web, http://www.w3.org/History.html

Wall A. (2006). Search Engine History, http://www.searchenginehistory.com/#before

World Intellectual Property Indicators (2011), WIPO Economics & Statistics Series, http://www.wipo.int/export/sites/www/freepublications/en/intproperty/941/wipo_pub_941_2011.pdf

Wu W., Yu C., Doan A. et Meng W. (2004). An interactive clustering-based approach to integrating source query interfaces on the deep web. In *Proceedings of the 2004 ACM SIGMOD international conference on Management of data* (pp. 95-106). ACM.

Yan J., Zhang B., Liu N., Yan S. Cheng Q., Fan W., ... et Chen Z. (2006). Effective and efficient dimensionality reduction for large-scale and streaming data preprocessing. *Knowledge and Data Engineering, IEEE Transactions on*, *18*(3), 320-333.

Yang Y., Zhang J. et Kisiel B. (2003). A scalability analysis of classifiers in text categorization. *In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 96–103.

Yu S., Wang C., Ren K. et Lou W. (2010). Achieving secure, scalable, and fine-grained data access control in cloud computing. In *INFOCOM, 2010 Proceedings IEEE* (pp. 1-9). IEEE.

Zimmer M. (2008). More on the 'Anonymity'of the Facebook Dataset–It's Harvard College'. *MichaelZimmer. org Blog*.

Zurfluh G., Chrisment C. et Pujolle G. (1993). Bases de données réparties». Bases de données Editions Techniques de l'ingénieur, Référence H3850.