3 Valorisation et intelligence des données massives

LOISIRS & CULTURE
ÉNERGIE, MOBILITÉ, NUMÉRIQUE
ENVIRONNEMENT, HABITAT, SANTÉ ET BIEN-ÊTRE, SÉCURITÉ
ALIMENTATION

➤ Correspond à une technologie clé 2015

MOTS CLÉS

Données massives, Big data, mégadonnées, analyse, stockage, données personnelles, prédiction, cloud computing, cybersécurité, capteurs, visualisation.

© Fotolia

Définition et périmètre

Définition

E-mails, réseaux sociaux, *smartphones* et objets connectés, capteurs et autres satellites génèrent un flot croissant d'informations hétérogènes et non-structurées. La valorisation et l'intelligence des données massives ou *Big Data* ou « mégadonnées »[1] désignent à la fois cette croissance exponentielle du volume des données disponibles sous forme numérique et la famille d'outils (technologies et algorithmes) permettant de les trier et les analyser en temps réel. De façon communément admise, le *Big Data* est défini par les 5V[2].

1. Volume : le *Big Data* fait référence en premier lieu au volume massif de données à traiter, augmentant à un rythme exponentiel. Il est en effet estimé que 90 % des données récoltées depuis le début de l'humanité ont été générées durant les deux dernières années[3] ;

2. Vélocité : la vélocité fait référence à la vitesse à laquelle l'information est créée, circule et est analysée. Notre tendance à dupliquer l'information sur plusieurs supports, à la partager sur différents objets ou l'effet viral des réseaux sociaux amplifient la vitesse de circulation des données. Par ailleurs, les outils du *Big Data* (ex : logiciels d'analyse) permettent de réaliser des études sur ces données en quelques heures, quelques minutes, voire en temps réel contre plusieurs jours auparavant ;

3. Variété : la variété fait référence à l'hétérogénéité des sources (capteurs, archives, réseaux sociaux, documents, applications mobiles, etc.) ainsi qu'à la diversité de formats des données, des informations classiques structurées dans une base jusqu'à des données non-structurées telles que le texte, l'email, la photo, la vidéo, et les métadonnées etc. Le *Big Data* permet de réunir toutes ces données et de les analyser ;

4. Véracité : l'un des enjeux du *Big Data* est d'améliorer la fiabilité des masses de données non-structurées, en améliorant la gestion du bruit et de la consistance, en organisant son accès et en y associant les algorithmes d'analyse correspondant aux besoins des utilisateurs ;

5. Valeur : caractéristique clé du *Big Data*, le volume massif de données n'a d'importance que s'il permet de générer du sens et donc de la valeur pour ces données. Le défi principal est donc d'identifier ce que les outils de valorisation des données massives peuvent apporter.

Les technologies associées à la valorisation et l'intelligence des données massives sont nombreuses et concernent à la fois les solutions *hardware*, *software* et les services associés.

On identifie d'une part les **technologies *software*** permettant d'optimiser les temps de traitement sur des bases de données massives[4] :

- Les bases de données NoSQL**, utilisant de nouveaux formats de stockage** (MongoDB, Cassandra ou Redis) implémentent des **systèmes de stockage** considérés comme plus performants que le traditionnel SQL pour l'analyse de données en masse ;
- Les infrastructures de serveurs permettent de réaliser le **traitement massivement parallèle**. L'infrastructure Spark, conçue spécifiquement pour les projets de *Big Data*, est aujourd'hui la plus utilisée pour traiter des données distribuées en *clusters* et exécuter plusieurs applications en simultané. Elle combine le système de fichiers distribué HDFS (Hadoop Distributed File System), la base NoSQL et l'algorithme MapReduce développé par Google ;
- **Le stockage des données en mémoire (Memtables)** accélère les temps de traitement de requêtes ;
- **Les technologies de *data mining*** permettent d'identifier l'information pertinente à l'aide d'outils statistiques perfectionnées (*clustering*, *machine learning*, data-viz, réseaux de neurones, algorithmes génétiques, etc.).

D'autre part, de **nouvelles plateformes *hardware* de serveurs** se développent pour s'adapter à la valorisation et l'intelligence des données massives. Aujourd'hui, la majorité des solutions *software* de *Big Data* fonctionne sur du matériel standard. Cependant, la plupart des acteurs considèrent que demain, la massification des données des entreprises nécessitera que les serveurs s'adaptent aux flux de plus en plus importants de données[5].

1 – Traduction officielle de la Commission générale de terminologie et de néologie en date du 22 août 2014. Cette étude fera néanmoins référence à l'anglicisme en raison de son utilisation dans le Plan de la Nouvelle France Industrielle

2 – Bernard Marr, *Big Data, using smart Big Data analytics to make better decisions and improve performance*, Broché, 2015

3 – *Big Data* Paris, *Le Guide du Big Data, Editions 2014/2015*, 2014

4 – ATKearney, *Big Data and the Creative Destruction of Today's Business Models*, 2013

5 – IBM, *The Evolution of Hardware and What It Means for Big Data*, 2013

Enfin, un large éventail de services s'est développé autour de la valorisation et de l'intelligence des données massives. Il s'agit notamment de *analytics as a service, infrastructure as a service, data as a service* et *business intelligence.*

Pourquoi cette technologie est-elle clé ?

La maîtrise et la collecte des données massives seront certainement l'enjeu majeur du XXIème siècle[6][7] et revêt donc pour la France un caractère stratégique. À l'heure où les données personnelles ont une valeur économique, détenir ces données et être capable de les analyser sera demain un critère de puissance mondiale. Dans différents secteurs métiers, la maîtrise des données sera à l'origine de profondes transformations des métiers et de l'organisation des entreprises (par exemple : la prévention de panne et la maintenance, le design de nouveaux produits etc.).

Un fort caractère stratégique

La valeur économique associée aux données personnelles et à leur exploitation par les entreprises est très élevée, voire pour certains, « illimitée »[8].

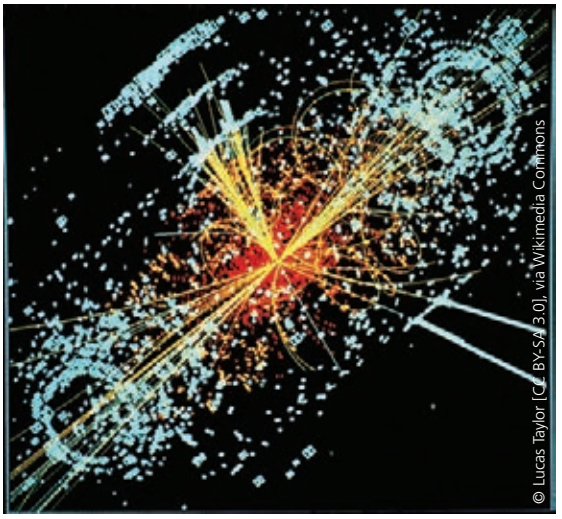
Les données personnelles, source primaire de cette « révolution » continueront à **augmenter de façon exponentielle** : réseaux sociaux, objets connectés, technologies mobiles et libéralisation des données publiques (*open data*) font exploser le volume des données disponibles. En 2020, il est estimé que 10,4 zettaoctets, soit 10 400 milliards de gigaoctets de données, seront partagés tous les mois sur Internet[9].

Compte tenu de la valeur associée à la détention des données et leur exploitation, l'appropriation des technologies de collecte et d'analyse des données massives par les entreprises françaises paraît aujourd'hui indispensable pour se maintenir dans la compétition mondiale.

Atouts de la France

Le positionnement des acteurs français sur cette technologie clé est d'autant plus essentiel que la France dispose **d'un système académique particulièrement performant** dans les disciplines sur lesquelles s'adossent la valorisation et l'intelligence des données massives. En matière de formation, l'ENSAE, référence dans le domaine de la statistique, propose une spécialisation en « Data Science », de même que Télécom ParisTech ou l'École Polytechnique qui propose depuis 2014 un mastère spécialisé en *Big Data*. Les offres de ce mastère bénéficient par ailleurs d'un mécénat des entreprises Keyrus, Orange et Thalès, preuve que les groupes français sont conscients de la nécessité de former et de renforcer les compétences des « *data scientists* ». Des organismes comme l'Inria, le CNRS et le CEA sont en outre à la pointe de la recherche et de l'innovation en «Data Science».

Par ailleurs, **un riche tissu de start-up** dans le domaine de la collecte et de l'analyse de données a émergé. À titre d'exemple, Dataiku, pionnier dans le *Big Data* en France, propose des solutions d'analyse prédictive tandis que CitizenData stocke, analyse et crée de la valeur à partir de données issues des capteurs sur textile.



À cela s'ajoute, enfin, **une volonté politique forte et affichée** de faire de la France l'une des références mondiales de la gestion des données massives. En 2011, un nouveau service du Premier Ministre a été créé : Etalab. Il s'agit d'une mission de création d'un portail des données publiques en ligne, permettant aux entreprises et acteurs publics de développer des

6 – « Création de l'Alliance *Big Data* », Site de CapDigital, 20/03/2013

7 – « *Big* data ; impact et attentes pour la normalisation », Livre blanc de l'AFNOR

8 – *Le Big Data parle. L'entendez-vous?,* Livre Blanc de l'EMC

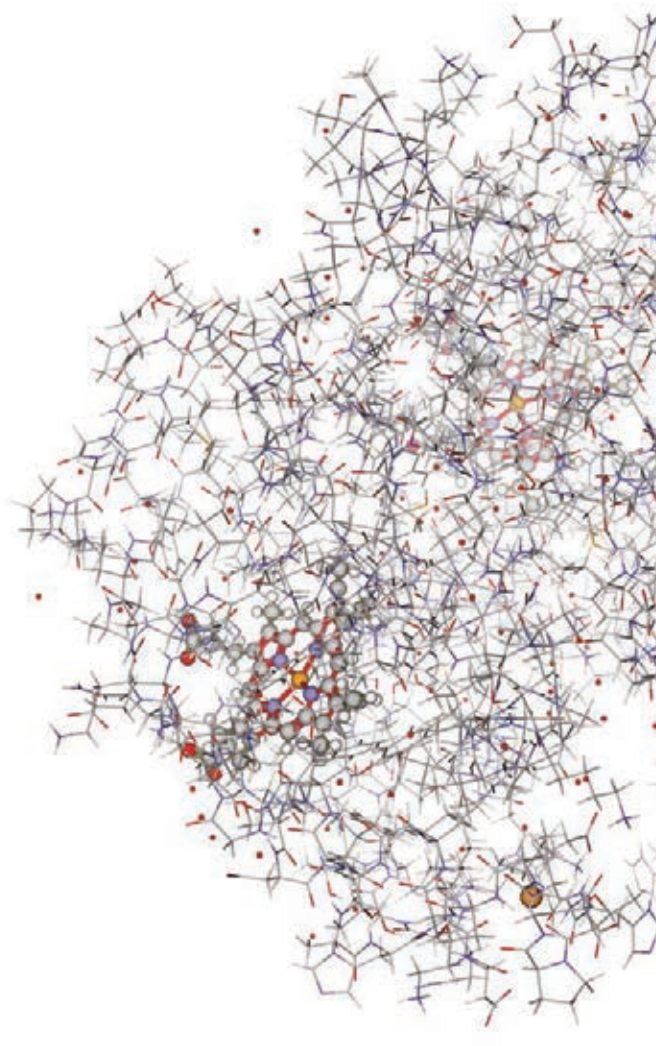
9 – « Vertigineux « *Big Data* », LeMonde.fr, 28/12/2012

nouveaux services à partir de ces données. En octobre 2013, le *Big Data* a fait l'objet d'un plan dédié de la Nouvelle France Industrielle, intégré depuis le 18 mai 2015 dans les neuf Solutions industrielles visant les marchés prioritaires pour la France. Plusieurs appels à projets ont déjà été lancés et seront publiés en 2015 et 2016 sur cette thématique dans le cadre du Programme d'Investissements d'Avenir et du Concours Mondial d'Innovation.

Liens avec d'autres technologies clés

La croissance du marché de la valorisation et de l'intelligence des données massives est étroitement liée à la maîtrise des technologies de production de données (objets connectés), de stockage (*Cloud* computing), de modélisation, de visualisation et de simulation, de calcul et d'analyse (analyse prédictive, analyse sémantique) :

- **L'Internet des objets (IoT)** ; l'explosion du marché de l'IoT multipliera la quantité de données personnelles et professionnelles disponibles et rendra d'autant plus nécessaire l'utilisation des technologies d'analyse des données pour les entreprises, les administrations ou encore les particuliers ;
- La maîtrise des **outils de modélisation, visualisation et simulation** est indispensable pour analyser et prendre des décisions à partir des données massives ;
- Le ***Cloud* Computing** : la valorisation et l'intelligence des données massives exigent une capacité matérielle hors du commun à la fois pour le stockage des données et pour les ressources nécessaires au traitement. Le *Cloud* permet l'exploitation de la puissance de calcul ou de stockage des serveurs informatiques distants par l'intermédiaire d'un réseau, généralement internet, offrant ainsi une capacité de valorisation et d'intelligence des données massives ;
- **Le calcul intensif** : l'émergence du *Big Data* et le développement des objets intelligents et connectés accroissent les besoins en calcul intensif. Ces technologies de calcul sont nécessaires à la valorisation et l'intelligence des données massives en ce qu'elles permettent d'analyser leurs flux[1] ;
- **Les technologies d'analyse prédictive** : les algorithmes prédictifs constituent une application directe des techniques de *Machine Learning* au *big data*. Par exemple, à partir d'un historique d'achats, de sessions de navigation sur un site internet, ces algorithmes peuvent prédirent les prochains besoins d'un consommateur ;

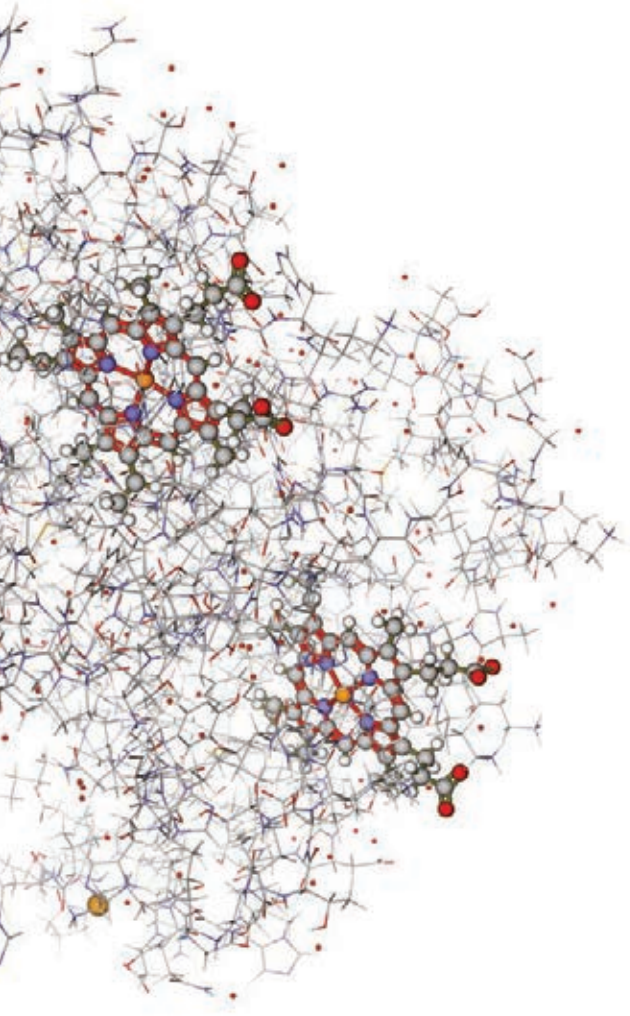


- **Les technologies sémantiques** ; ces technologies permettent d'identifier les données pertinentes et signaux faibles cachés dans les données massives. Les possibilités d'extraction vont de la simple reconnaissance de personnes ou d'entreprises à l'analyse d'opinion en passant par la catégorisation thématique ;
- **La cybersécurité** : l'augmentation du volume de données personnelles disponibles sur internet néces-

1 – CNRS, *Livre blanc du calcul intensif*, 2012

site de mettre en place des technologies de cybersécurité plus performantes pour protéger tant les données des utilisateurs (organisations et individus) que la possibilité d'intrusion et de prise de contrôle de systèmes, comme par exemple des objets connectés ;

■ **Les infrastructures de 5ème génération** : le traitement de données de plus en plus importantes en



© Fotolia

un temps restreint nécessitera la mise en place d'un réseau plus rapide. Le développement de la 5G permettra donc, conjointement, le développement de solutions de traitement de Big Data tant centralisées que distribuées ;

■ **L'intelligence artificielle (IA) et le *deep learning*** : les algorithmes de *deep learning* s'inspirent de ceux de l'intelligence artificielle pour extraire automatiquement des informations du *Big Data* ou des représentations de données avec un niveau élevé d'abstraction.

Les marchés

Aujourd'hui, toutes les projections du marché du *Big Data* prévoient **une très forte croissance d'ici 2020**. Elles concernent les solutions de serveurs, stockage, réseaux et logiciels (bases de données relationnelles NoSQL, Hadoop...) ainsi que les services associés.

Un marché en très forte croissance

Le chiffre d'affaires du marché de la valorisation et de l'intelligence des données massives (hors services) devrait croître de **40 % par an** pour atteindre **44,4 milliards d'euros en 2018**[2]**.** En intégrant non seulement les logiciels mais aussi les services, le marché est évalué à 31 milliards d'euros en 2013 et à **105 milliards d'euros en 2018**, soit une progression annuelle **de 29,6 %**[3]**.** Les leaders du marché mondial sont principalement américains. Il s'agit notamment de Google, Amazon, Facebook, Apple, IBM, Intel, Microsoft, TeraData, *Cloud*era, Oracle, EMC, Hortoworks et DataMeer.

En France, le marché de la valorisation et de l'intelligence des données massives est évalué à seulement 387 millions d'euros en 2013 mais est considéré comme particulièrement prometteur avec un taux de croissance de **40 % par an** (logiciel et services)[4]. Selon les estimations du plan dédié de la Nouvelle France Industrielle, intégré dans la Solution industrielle « Économie des données », le marché en France devrait atteindre **9 milliards d'euros en 2020**[5]. Les leaders du marché français sont notamment Atos, Thales, Criteo, Orange et Dataiku.

Des applications dans tous les secteurs

L'émergence des technologies de la valorisation et de l'intelligence des données massives décuple les possibilités d'analyse dans tous les secteurs comme l'illustre le tableau ci-dessous :

---

2 – Données de *Transparency Market Research*

3 – Données d'*ABI Research*

4 – « Le marché du «*Big Data*», nouveau graal de l'informatique », LeFigaro.fr, Tech & Web, 02/04/2014

5 – Ministère de l'Economie, de l'Industrie et du numérique, *Les 34 plans de la nouvelle France industrielle*, 2013

Secteurs	Exemple de projets
Santé	Projet de recherche en génomique mené par France génomique
Marketing, CRM, publicité	Solution de re-ciblage publicitaire utilisant des algorithmes de *machine learning* pour construire les bannières qui correspondent aux attentes des utilisateurs
Environnement	Projet « Dada » du CNRS : analyse de l'évolution climatique au niveau mondial
Lutte contre la fraude	Projet « Brand WatchDog » mené par l'entreprise Data & Data
Journalisme	Projet « The Migrant Files » mené par l'entreprise Journalism+++
Banque et assurance	Projet d'assurance évolutive mené par la société Progressive avec le lancement du service « *Pay as you drive* »
Loisirs	Projet « X-Field Paintball » de la société PCB Team : création d'une base de données géantes sur la communauté du paintball
Énergie	Projet « Deepky » de Cofely Service utilisant des algorithmes de data-analytics
Mobilité	Création d'un algorithme d'analyse des données de vol. Projet mené par l'entreprise Safetyline

Sources : Atelier d'experts « Communication Numérique » du 31 mars 2015, organisé dans le cadre de l'Étude Technologies clés 2020 et Guide du Big Data 2014/2015
Tableau non-exhaustif des secteurs d'application du *Big Data* et des projets en cours

Le marché de la valorisation et de l'intelligence des données massives bénéficie d'un potentiel très important pour les entreprises et administrations publiques[6] :

- Le potentiel du *Big Data* dans le secteur de la santé serait de **275 milliards d'euros** ;
- Les économies potentielles liées à la collecte et au traitement des données massives pour les administrations publiques en Europe sont estimées à **250 milliards d'euros** ;
- La valeur des données de géolocalisation pour les prestataires de services représenterait **92 milliards d'euros** ;
- Enfin, la valeur des données personnelles en Europe est estimée **à 315 milliards d'euros**.

Les défis technologiques à relever

Le stockage des données massives

Les entreprises de taille moyenne détiennent toutes un centre de données (*data centers*) ou externalisent cette fonction de stockage et d'archivage à des prestataires spécialisés. La croissance exponentielle de la masse des données des entreprises interroge sur le futur du stockage de données. L'augmentation du volume des centres de données ne pourra en effet pas suivre la courbe exponentielle de croissance du volume de données générées dans le monde.

Le traitement et la qualité des données produites

Les technologies de traitement et d'analyse des données ne garantissent pas aujourd'hui une fiabilité totale des données analysées. En effet, le traitement de grands volumes de données peut accroître la marge d'erreur si les données ne sont pas intègres à la base. Pour contrer cela, de nouvelles solutions sont développées pour mieux percevoir la source de la donnée et réduire le taux d'erreur. Des fournisseurs spécialisés dans l'analyse et le nettoyage de la donnée externe ont également vu le jour. Mais l'analyse humaine reste quoi qu'il en soit indispensable. La **montée en compétence** et la **formation de *data scientists*** présentent à ce titre un intérêt stratégique pour l'entreprise.

Le temps réel

Un autre enjeu technologique de la valorisation et de l'intelligence des données massives est la quête du temps réel *via* la réduction du temps de traitement. Les nombreux travaux de recherche technologique en cours proposent des solutions différenciées pour accélérer le traitement des données, à l'instar d'*in-memory*[7]. Le développement des infrastructures de 5ème génération pourrait par ailleurs augmenter considérablement la vitesse de traitement des données.

6 – « Infographie ; *Big Data*, un marché à 100 milliards de dollars », Usine-digitale.fr, 08/04/2014

7 – *Big Data* Paris, *Le Guide du Big Data, Editions 2014/2015*, 2014

Sécurité et anonymisation des données

Face à un volume toujours plus massif de données disponibles, **l'enjeu de leur anonymisation est devenu un enjeu clé.** La nouvelle gestion des données du *Big Data* tend en effet vers ces procédés d'anonymisation des données qui garantissent aux individus une préservation de leur vie privée, tout en permettant aux entreprises de valoriser les informations contenues dans les données personnelles qui, une fois dissociées d'une personne identifiable, ne sont plus soumises à la loi Informatique et Libertés. L'objectif est ainsi d'aboutir à des procédés d'anonymisation irréversibles et absolus, rendant impossible toute identification, alors même que les recoupements massifs permis par le *Big Data* apparaissent antagonistes avec cette recherche. Des critères qui garantissent un niveau d'anonymisation suffisant sont encore à affiner. Cet enjeu technologique est en ce sens très lié au défi de confiance entre opérateurs et clients.

Les défis commerciaux et d'usage à relever

La croissance exponentielle des données dont disposent les entreprises créent des opportunités nouvelles. En améliorant la connaissance des comportements et préférences de consommation de leurs clients, ainsi que leurs processus de production, les entreprises peuvent proposer de nouveaux produits ou améliorer l'existant, personnaliser davantage leurs offres pour mieux répondre aux besoins de leurs clients et les fidéliser, et *in fine* améliorer leurs résultats commerciaux.

Le « défi de la confiance » entre opérateurs et clients

L'utilisation de données personnelles créée un **risque de réputation fort** pour l'entreprise positionnée sur ce marché, si ses clients n'ont pas confiance dans la manière dont leurs données sont protégées et utilisées[8]. La confiance est centrale dans la relation qui lie l'entreprise avec son client et est d'autant plus importante dans le cas de l'utilisation de données personnelles. Les évènements récents qui ont marqué l'opinion publique - tels que la surveillance exercée par la NSA ou l'attaque informatique de Sony qui a révélé au public l'exploitation d'un nombre important de données personnelles - ont mis en évidence et accentué la sensibilité du public à la protection de ces données et induit le besoin d'être rassuré. Selon une étude récente[9], en moyenne, 78 % des personnes interrogées en Grande Bretagne, Allemagne, France, Italie et Espagne considèrent qu'elles doivent être prudentes lorsqu'elles partagent des données personnelles en ligne.

Une sous-estimation de la valeur et de l'utilité du Big Data par les entreprises françaises

La valorisation des données massives doit « révolutionner » le travail des entreprises mais encore faut-il qu'elles le réalisent. D'après une étude publiée en juin 2015, les entreprises françaises sont particulièrement en retard dans le domaine et n'ont pas pris conscience de la valeur que peut leur apporter la mise en place d'une stratégie de valorisation et d'intelligence des données massives. Pour deux tiers d'entre elles, le *Big Data* est un concept «intéressant, mais trop vague pour constituer un levier de croissance»[10].

Un retard dans l'intégration du Big Data dans les entreprises françaises

En 2015, seules 18 % des entreprises françaises ont des plans d'actions en cours de déploiement dans le *Big Data*, et seules 17 % d'entre elles sont « très matures » dans l'exploitation de leurs données clients. La collecte de données est limitée aux canaux traditionnels. Les données non-structurées sont insuffisamment analysées. Les entreprises manquent de compétences analytiques pour traiter leurs données client ainsi que d'outils spécifiques pour les données non-structurées.

Selon l'Observatoire de l'Innovation de l'Institut de l'entreprise[11], cet état de fait peut s'expliquer par plusieurs limitations qui trouvent leur source dans les entreprises elles-mêmes :

- Des difficultés de coordination entre les différents services de l'entreprise, impliquant souvent la coûteuse mise en place d'un département de *data scientists* ;
- La complexité de l'évaluation des bénéfices en termes de productivité et de croissance des straté-

8 – Boston Consulting Group et DLA Piper, *Le Big Data face au défi de la confiance*, juin 2014

9 – *Ibid*

10 – EY, *(Big) data ; où en sont les entreprises françaises ?*, 2014

11 – Observatoire de l'Innovation de l'Institut de l'entreprise, *Faire entrer la France dans la troisième révolution industrielle ; le pari de l'innovation #1 Big Data*, mai 2014

gies liées au *Big Data*, face à des besoins d'investissements importants ;

- La mise à niveau des compétences en statistique, informatique et management pour nombre de cadres.

A ces défis d'intégration et d'appropriation de la valorisation et de l'intelligence des données massives par les entreprises françaises s'ajoute un enjeu majeur pour l'industrie du *Big Data*, celui de la protection des données personnelles ou organisationnelles.

Les enjeux réglementaires

L'exploitation par des entreprises de données personnelles à des fins commerciales (base de données marketing, ciblage publicitaire, etc.) pose inévitablement la question de la protection des données personnelles. Qui peut collecter ces données ? Qui en détient la propriété ?

En France, le traitement de données à caractère personnel est régi par les dispositions de la loi « Informatique et Libertés » du 6 janvier 1978. Cette loi définit une donnée personnelle comme « toute donnée permettant d'identifier directement ou indirectement une personne physique » et énonce les principes relatifs à la protection des données :

- Finalité et proportionnalité de la collecte des données ;
- Pertinence des données traitées ;
- Conservation limitée des données ;
- Sécurité et confidentialité ;
- Respect des droits des intéressés : loyauté et transparence.

Ces principes constituent autant de défis du point de vue de la conformité règlementaire des opérations de valorisation et d'intelligence des données massives.

La finalité des opérations de *Big Data* est souvent imprécise, la proportionnalité et la pertinence sont donc difficiles à délimiter[12]. Les opérateurs recherchent des signaux faibles dans la masse de données et des corrélations, mais ne savent souvent pas sur quoi ils vont s'arrêter : c'est le principe de « sérendipité »[13]. Du fait des rapprochements, croisements et analyses de données issues de sources diverses et de la dispersion des moyens de traitement, la collecte et l'utilisation de ces données massives doivent être précisément encadrées sur le plan juridique.

Par ailleurs, **la sécurité et la confidentialité des données** sont capitales pour les opérateurs du *Big Data* étant donné la valeur potentielle qu'une analyse efficace du patrimoine informationnel de l'entreprise et du particulier peut générer. En stockant des données stratégiques sur lesquelles elle compte appuyer ses décisions, l'entreprise s'expose à des phénomènes de cybercriminalité et de piratage. Ce fut le cas d'*Ebay* en 2014, victime d'un vol important de données de ses clients.

Les solutions de stockage sur serveurs et dans le *Cloud* doivent de ce fait répondre à ce risque d'insécurité sur les réseaux. Des actions sont en cours dans le cadre de la Solution industrielle « Économie des données » avec la création d'un label sur la sécurité dans les services de cloud computing (Label Secure Cloud de l'ANSSI) qui permettra aux entreprises et acteurs publics d'avoir confiance en ces nouveaux services.

Corollaire de la sécurité des données, **le principe de collecte loyale impose que les intéressés consentent au traitement de leurs données personnelles.** Le 13 mars 2014, le Conseil d'État dans son arrêté « PagesJaunes » a donné raison à la CNIL à propos de l'obligation d'informer les internautes sur la collecte d'informations issues du *web* social.

Face à l'absence de cadre juridique spécifique, l'Union européenne travaille depuis 2012 sur un projet de règlement européen unifiant le droit de tous les États en la matière. L'objectif principal de ce règlement est d'établir clairement la finalité et les conditions d'utilisation des données personnelles[16]. Ce nouveau règlement devrait être proposé d'ici la fin de l'année 2015. En France, le projet de loi numérique en cours de rédaction devrait aussi aborder certaines thématiques liées aux données personnelles, comme la portabilité.

---

12 – Société Française de Statistique, *Enjeux Ethiques du « Big Data » : Opportunités et risques*, Séminaire organisé par le groupe « Statistique et enjeux publics » de la SFdS le 22 mai 2014

13 – «Protection des données personnelles et *Big Data* ; inconciliables, vraiment ?», Site de Silicon.fr, 20/04/2014

14 – Données du *Ministère de l'économie et des finances*

15 – GAFAM : acronyme témoignant de l'hégémonie de cinq acteurs américains sur le marché du Big Data. Il s'agit des initiales des entreprises : Google, Apple, Facebook, Amazon et Microsoft

16 – Commissariat Général à la stratégie et la Prospective, Marie-Pierre Hamel et David Marguerit, *Analyse des Big Data Quels usages, quels défis ?* », novembre 2013

Analyse AFOM

ATOUTS

Des instituts d'enseignement supérieurs de renommée internationale : l'Institut Mines-Télécom (Télécom Paristech notamment), École Centrale Paris, l'ENSAE, l'École Polytechnique, l'École normale supérieure (Cachan), l'École normale supérieure (Ulm), etc.

Des ressources d'informaticiens et mathématiciens plébiscités dans le monde entier sur des sujets clés comme le *Big Data* ou l'intelligence artificielle

Des centres de recherche à la pointe sur le *Big Data* tels que le CNRS, le CEA List et l'INRIA

Un dynamisme de l'action publique à travers :

L'intégration d'actions *Big Data* dans la Solution industrielle « Économie des données » de la Nouvelle France industrielle

La création en 2011 de la mission Etalab, « portail unique interministériel des données publiques »

La Présidence française du « Partenariat pour le gouvernement ouvert » (*Open Government Partnership*) en 2016[1]

La création de la conférence *Big Data* Paris

FAIBLESSES

Un manque de compétences analytiques et notamment de ressources de *data scientists* pour subvenir aux besoins de ces prochaines années

Une faible perception des entreprises et des particuliers de la valeur des *Big Data* dans la prise de décision stratégique

OPPORTUNITÉS

Un marché mondial estimé à plus de 40 milliards d'euros en 2018

Un marché français estimé à 9 milliards d'euros en 2020

Un potentiel de 137 000 emplois d'ici 2020 en France

Une nouvelle réglementation européenne sur la protection des données à caractère personnel

MENACES

Un monopole des GAFAM[2] détenant les plateformes globales d'échange et collecte des données massives

Une faible protection des données personnelles

Une atteinte à la sécurité économique des entreprises à travers la vulnérabilité du patrimoine informationnel de l'entreprise

Facteurs clés de succès et recommandations

Aux pouvoirs publics

■ Intégrer le traitement des données massives dans l'action publique : en tant que grand opérateur de données, l'État doit être un acteur phare du dispositif et se montrer exemplaire en la matière ;

■ Participer à l'établissement d'un cadre réglementaire favorable à l'émergence d'une industrie française de valorisation et d'intelligence des données massives afin de répondre aux enjeux économiques et de souveraineté.

Aux entreprises

Mettre en place une stratégie efficace de valorisation et d'intelligence des données massives. Pour ce faire, quatre facteurs clés de succès peuvent être mentionnés :

■ Impliquer fortement la direction générale de l'entreprise ;

■ Définir la stratégie en impliquant l'ensemble des directions et en travaillant sur leur transversalité ;

■ Recruter et former des *data-scientists ;*

■ Définir une feuille de route agile et un plan d'action concret ;

■ Garantir la sécurité des données pour créer un climat de confiance ;

■ Expérimenter sur des problématiques *Big Data* concrètes et précises.

Aux académiques

■ Améliorer les compétences en *analytics* à travers la formation de *data scientists ;*

■ Accroître l'offre de formations continues, considérée par les experts comme plus efficaces pour former des *data scientists* à horizon 2020.

Acteurs clés

Organismes de recherche et de formation

En France, **l'INRIA, le CEA List, et le CNRS** sont les principaux centres de recherche travaillant sur la valorisation et l'intelligence des données massives. **L'IRT System X** est un acteur important qui mène des travaux de recherche technologique dans ce domaine. Par ailleurs, les principales grandes écoles françaises ont également lancé des programmes de recherche et formation sur le *Big Data*. **L'Institut Mines-Télécom** propose un programme complet de recherche et d'enseignement pluridisciplinaire sur le *Big Data* accessible en formation initiale et continue et concernant 13 enseignants chercheurs, 50 doctorants et une centaine de diplômés par an. L'école **Télécom ParisTech** a lancé à la rentrée 2013 un nouveau mastère spécialisé «*Big Data* : Gestion et analyse de données massives».

L'École Polytechnique, l'ENSAE et l'École Centrale Paris ont également mis en place des programmes spécialisés sur la thématique[17]. À titre d'exemple, l'École Polytechnique propose le « *Data Sciences Starter Program* », l'École Centrale Paris a créé un programme de formation continue à destination des cadres dirigeants, chefs de projets, managers des systèmes d'information et experts sur la thématique « *Big Data* - Enjeux et opportunités » ; enfin, l'ENSAE a structuré un programme de spécialisation en *Data Science*, visant à former des *data scientists,* et ce tant en formation initiale que continue.

Grands groupes

La France est encore trop absente des « couches basses »[18] du marché c'est-à-dire au niveau des technologies de base et des infrastructures. Ces « couches » sont aujourd'hui quasiment exclusivement occupées par des acteurs américains tels que Google (création de l'architecture *Hadoop*) ou encore Amazon et Cloudera. Les Français ont pris le train en marche, mais sont présents principalement au niveau des applications. **Atos et Bull, Thales, Orange et Keyrus** sont considérés comme les principaux groupes français de la valorisation et de l'intelligence des données massives. Le groupe Technicolor, leader mondial du secteur des médias et des divertissements, s'est lancé dans le

17 – Pour en savoir plus : sites Internet de Polytechnique, de l'École Centrale Paris et de l'ENSAE

18 – « L'équipe de France du *Big Data* ». LesEchos.fr, 15/10/2013

Big Data avec la création en 2014 de Virdata, en partenariat avec IBM. Ce service de *cloud* pour l'Internet des objets comprend des technologies avancées de gestion et analyse des *Big Data*.

Entreprises de taille intermédiaire (ETI)

Criteo est l'ETI leader dans le domaine de la valorisation et de l'intelligence des données massives en France. L'entreprise a mis en place une architecture informatique de pointe dans le domaine du *Big Data* et un algorithme capable de prédire les intentions d'achat des internautes à partir de leur historique de navigation afin de mieux cibler les publicités affichées.

Start-up et PME

L'écosystème du *Big Data* en France est majoritairement dominé par les très nombreuses start-up spécialisées dans l'*analytics*, la sémantique et les modèles prédictifs. On peut citer, parmi d'autres, **CitizenData**, solution hébergée de collecte, stockage et analyse de données issues de capteurs, ou encore **Dataiku**, qui édite une solution d'analyse de données et de construction d'applications prédictives[19], **Mesagraph**, qui mesure l'audience sociale des programmes des chaînes de télévision en analysant des millions de tweets par mois et enfin, **Syllabs**, qui a récemment développé une offre de solutions d'analyse sémantique dédiées pour traiter les données massives dans les domaines de l'e-commerce, l'e-tourisme et les media.

Organismes de soutien et d'interface

Enfin, ces entreprises sont soutenues et accompagnées par les pôles de compétitivité français **Cap Digital, Images & Réseaux, SCS et Systematic**. Cap Digital a participé à la création de l'Alliance *Big Data* en 2013 dont l'objectif est de « contribuer à la construction d'une vision commune et de favoriser le développement de nouveaux services et projets dans le domaine du *Big Data* en France »[20].

Avec Cap Digital, quatre autres animateurs ont été sélectionnés pour lancer des appels à projets « Challenges Big Data », permettant de mettre en relation grands groupes et start-up. Il s'agit notamment de Numa, Images & Réseaux, TUBA Lyon et Euratechnologies.

Position des acteurs français

Position des entreprises françaises dans la compétition mondiale	
En position de leadership	
Dans la moyenne	●
En retard	

Les acteurs français sont en retard dans la compétition mondiale par rapport aux États-Unis qui dominent très largement le marché[21]. La majorité des leaders mondiaux de fourniture de solutions de *Big Data* sont en effet américains. À l'échelle européenne en revanche, la France se positionne au même rang que l'Allemagne, devant le Royaume-Uni[22].

Position des acteurs académiques français dans la compétition mondiale	
En position de leadership	●
Dans la moyenne	
En retard	

Les grandes écoles de statistiques et de mathématiques françaises - tels que Télécom ParisTech, l'ENSAE, l'École Normale Supérieure de Cachan, l'École Normale Supérieure Ulm, l'École Centrale ou l'École Polytechnique – et les acteurs de la recherche publique (CNRS, INRIA, CEA etc.) ont permis à la France de développer un système académique performant, de former des talents et de se positionner en leader dans les disciplines de la valorisation et de l'intelligence des données massives. De nombreuses structures travaillent également sur ce domaine comme par exemple les LabEx SMP (porté par la Fondation Sciences Mathématiques de Paris), Digicosme et Ecodec, Les EquiEx CASD et Digiscope ou bien encore les Lidex CDS, ISN, etc.

19 – La start-up a levé 3 millions d'euros début 2015

20 – Site Internet de Cap Digital

21 – Ateliers d'experts réalisés dans le cadre de l'Étude Technologies Clés 2020

22 – Teradata, Communiqué de Presse : « *Les entreprises françaises exploitent davantage le « nouveau » Big Data que leurs homologues anglaises* », septembre 2014