La normalisation et le Big Data

Par Charles HUOT
People in the Sun

Introduction

Les perspectives économiques, industrielles, techniques et sociétales associées à la collecte et l'exploitation des Big Data représentent d'importantes opportunités, mais aussi des risques qu'il convient d'appréhender, d'anticiper et de maîtriser. L'une des caractéristiques des Big Data est qu'elles concernent tous les secteurs et qu'elles interfèrent avec tous les domaines qui connaissent des situations de rupture technologique comme la ville intelligente, la e-santé, les moyens de production, les réseaux intelligents, les objets connectés, etc. Les technologies classiques de gestion des données, basées depuis des années sur des mécanismes transactionnels à partir du stockage et de l'interrogation de données structurées, ne suffisent plus car elles présentent des limites avec l'avènement de l'Internet social (*social web*), de la mobilité, des « *smart* »*phones* et des tablettes, des capteurs et des objets connectés. De nouvelles technologies prennent en compte les caractéristiques des Big Data et permettent leur analyse en temps réel ou quasi réel.

Développer l'exploitation des Big Data demande des leviers. Les normes sont des outils volontaires venant en appui des entreprises pour apporter des solutions aux enjeux des Big Data. Elles sont un facteur d'ouverture des marchés et de confiance entre partenaires. Leur développement, comme leur adoption, est un enjeu concurrentiel. Dans un environnement national, européen et international de plus en plus complexe et multiforme, il importe que les acteurs prennent conscience de l'importance de ce levier et réfléchissent à une stratégie appropriée dans le domaine clé des Big Data pour en faire des outils efficaces au service des différentes parties intéressées.

Par Big Data, nous entendons l'exploitation de grandes masses d'informations (téraoctets) composées de données souvent hétérogènes (multimédia, capteurs, réseaux sociaux, téléphonie, etc.) avec des outils extrêmement véloces (permettant d'envisager des décisions en temps réel), ce qui implique le cas échéant des moyens non conventionnels (par exemple, une base de données NO SQL). Le présent article analyse les enjeux pour la normalisation associés à la mise en œuvre de processus de collecte, traitement et exploitation de grandes masses de données, souvent hétérogènes et non structurées, par exemple en provenance d'Internet, de réseaux sociaux publics ou d'entreprises, de réseaux de communication, de capteurs associés à des réseaux intelligents, des objets mobiles, des dispositifs de sécurité ou des sites de production industrielle, etc. Pour ce qui est des enjeux normatifs, on ne se limite pas aux questions technologiques, bien qu'elles soient importantes, mais on considère l'outil normatif dans sa capacité à proposer aux entreprises et aux acteurs des interfaces, des pratiques et des modes d'organisation partagés et adaptés aux changements profonds qu'imposent les modèles économiques des Big Data.

Enjeux, perspectives et opportunités

L'exploitation des Big Data bénéficie de plusieurs ruptures technologiques rendant possibles l'intégration, le traitement, l'interprétation et la représentation de données hétérogènes provenant de différentes sources. Ceci permet de placer désormais les données au cœur des modèles économiques et d'apporter aux organisations une forte valeur ajoutée par un meilleur ciblage de leurs services à travers à la fois une meilleure connaissance de leur environnement et une optimisation

de leurs processus. Une convergence entre les domaines d'affaires est ainsi possible et ceci marque sans doute le début de nouveaux modèles économiques qui redéfinissent les relations entre les producteurs, les distributeurs et les consommateurs ou les biens et services. Ceci étant, la complexité liée à ces relations s'est considérablement accrue, rendant la prise de décision extrêmement difficile pour une organisation donnée. La nouveauté dans les Big Data réside dans le besoin d'exploiter de gigantesques volumes de données liés à la diversité et à la multiplicité des sources qui sont désormais accessibles.

Les freins

Le contrôle sur les données

Les entreprises sont conscientes de l'importance des données qu'elles détiennent. Le caractère stratégique peut imposer que ces données ne puissent être partageables. Ceci limite intrinsèquement leur exploitation dans un contexte de Big Data du fait que celui-ci accroît le risque de fuite d'informations. Les organisations ont une sensibilité renforcée aux événements et aux incidents de sécurité, puisque les dispositifs d'exploitation de Big Data imposent des systèmes très intégrés et font appel à des prestataires spécialisés. Lorsqu'elles sont partageables, les données posent la question des licences d'exploitation et de la propriété intellectuelle associée, ainsi que celle de la traçabilité de leur usage. Les notions de droits d'usage des données conduisent aussi à des questions de non-répudiation. L'enjeu de traçabilité des données et des traitements est d'autant plus important dans un contexte de données ouvertes. Ceci concerne en particulier l'exploitation de données culturelles, celle des données de la recherche, et induit celle de l'identification des auteurs ou/et des chercheurs. Il existe déjà des pratiques, des règles et des normes, mais dans des domaines limités que le Big Data bouleverse.

La sécurité juridique et l'éthique

La manipulation de données à caractère personnel conduit à des enjeux sur les processus d'anonymisation et soulève la question de leur cryptage non réversible. Par ailleurs, le risque d'image peut être important pour certains acteurs (collectivités publiques, organismes financiers, etc.) en raison des effets et des dérives « Big Brother » liés à la collecte massive de données. Ce risque vaut aussi pour les grandes entreprises qui souhaitent maîtriser leur réputation sur les réseaux sociaux et à travers celle-ci leur relation client. Les enjeux sociétaux peuvent aussi se traduire par des contraintes sur les processus de présentation de l'information et leur usage.

Pour les entreprises, mettre en œuvre une démarche de Big Data représente un enjeu de compétitivité avec la possibilité de générer de nouveaux profits et de se positionner dans de nouvelles activités. Pour les acteurs publics, c'est disposer d'une capacité d'optimiser leur fonctionnement et de proposer de nouveaux services aux citoyens. Pour les citoyens, c'est une possibilité d'être acteurs conscients dans l'écosystème des Big Data, et de bénéficier de nouveaux services qui amélioreront leur qualité de vie. Pour bénéficier de ce nouveau modèle économique, il est nécessaire de décloisonner les silos et de jouer sur la transversalité entre les métiers, ce qui impose une démarche de transformation de l'entreprise. Il s'agit en effet d'être en mesure d'appréhender l'information mais aussi de prendre conscience du patrimoine informationnel au sein des entreprises et des organisations. Ceci suppose une compréhension et une maîtrise de la valeur des données, de ce qui est partageable et de ce qui ne l'est pas, dans quelles conditions, avec des enjeux en termes de propriété intellectuelle, de connaissance de la réglementation notamment en matière de données à caractère personnel.

Pour les acteurs publics, les données deviennent essentielles pour le pilotage des territoires. La donnée se trouve au cœur de nombreux concepts émergents (Open Data, Big Data).

Les entreprises comme les collectivités territoriales sont cependant confrontées à des questions techniques :

- En premier lieu, la qualité de l'information résultant d'une démarche Big Data est directement liée à la qualité des jeux de données en entrée. La qualité des données est un enjeu important en raison de l'exploitation de données de différentes sources souvent non homogènes, ce qui a des répercussions sur les processus de traitement analytique et sémantique.
- En second lieu, il est utile de pouvoir mettre en œuvre une interopérabilité qui se traduit dans les processus de capture des données, dans la mise en œuvre de référentiels de métadonnées, dans les processus de filtrage et d'extraction de l'information ainsi qu'au niveau de la restitution des résultats.
- Enfin, il faut assurer une sécurité des données afin d'en garantir l'intégrité et la confidentialité. Il s'agira notamment de sécuriser l'intégralité du flux de traitement des données, et de disposer de mécanismes de non-répudiation (à la source) et d'approbation (du résultat).

Tout ceci nécessite un apport de compétences spécifiques (analystes, statisticiens, juristes de la donnée…), mais aussi des infrastructures et des technologies particulières.

Conclusion

Pour que l'écosystème du Big Data et de l'Intelligence Artificielle se développe, notamment sur le plan des modèles économiques, il convient de prendre en compte des problématiques-clés relatives aux données :

- L'interopérabilité des données et des systèmes d'information est essentielle pour permettre une utilisation de la masse d'information disponible, dans une démarche de pilotage global.
- La sécurisation des données doit être prise en compte dans les spécifications techniques des systèmes de collecte, de stockage et de traitement de l'information. Elle concerne aussi le niveau juridique pour ce qui est de la propriété des données et des résultats ainsi que l'exploitation des licences d'exploitation.
- Le respect des contraintes relatives aux données personnelles, encadrées par la CNIL, doit être pensé en amont de l'implémentation du système. Une anonymisation des données est l'une des techniques-clés pour faciliter leur exploitation (cf. encadré).
- Les conditions de gestion (mutualisation SIG…) et de transmission des données, entre entités privées, publiques ou publiques et privées doivent être précisées par une organisation spécifique.

Norme volontaire sur la métrologie de réidentification de jeux de données anonymisés : exemple dans le domaine de la santé

L'exploitation des Big Data dans le domaine touchant à la santé comporte la manipulation et le traitement de données particulièrement sensibles. C'est pourquoi le CNAM (laboratoire CEDRIC) a mis au point des modèles de mesure (quantification) du risque de réidentification d'un jeu de données, particulièrement en présence des croisements de jeux de données prévus et des traitements statistiques associés, sous les mêmes hypothèses concernant l'échantillon de données que celles de Latanya Sweeney (2002). La méthode mise au point permet d'estimer une probabilité de réidentification dans un temps raisonnable et présente l'avantage d'être rapide et originale.

Le CNAM a sollicité l'AFNOR afin d'établir un document de référence fondé sur ces travaux. À terme, l'outil développé pourrait permettre de quantifier un préjudice tel qu'un vol de données.

Dans le contexte de financiarisation et de judiciarisation des données dangereuses, la méthode pourrait permettre à des juges de prendre une décision de façon plus objective.
Si le numérique offre de nombreuses opportunités dans le domaine de la santé, il est également porteur de risques, principalement liés à la gouvernance des données et à l'exploitation de ces données particulièrement sensibles. En particulier, le nouveau Règlement général sur la protection des données (RGPD) s'applique puisque ces données concernent des personnes. Par ailleurs, en France, comme dans le reste de l'Europe, l'accès aux données de santé relève de la réglementation. La réglementation sur les données de santé définit dans quels cas les données de santé peuvent être utilisées. La règle générale est que les données de santé ne sont pas utilisables sauf exceptions.

Pour prévenir ces risques, l'une des méthodes couramment utilisées dans les processus Big Data et plus particulièrement en santé consiste à rendre anonymes (« anonymiser ») les données pour leur ôter tout caractère personnel.

Pour ce faire, différentes techniques existent dont certaines font l'objet de normes. Malgré tout, force est de constater que le risque de « data breach » de données anonymisées, et donc de réidentification des personnes, est important :

- 72,7 % des attaques de réidentification ont eu lieu depuis 2009 (tous secteurs confondus).
- Données utilisées : essentiellement des *data sets* de type ADS (gros tableau Excel + graphes)

Disposer d'une méthode pour évaluer le risque de réidentification inhérent au processus d'anonymisation représente par conséquent un enjeu important. Or, les techniques couramment utilisées pour évaluer la robustesse de bases de données ayant été anonymisées font appel à la « force brute », et restent de ce fait particulièrement coûteuses en temps et en ressources de calcul. De ce fait, elles ne sont employées qu'avec parcimonie.

À ce jour, il n'existe pas à notre connaissance de norme concernant l'évaluation du risque de réidentification de données anonymisées. L'enjeu d'une démarche de normalisation dans ce domaine est de valoriser la méthode mise au point par le CNAM afin de disposer d'un état de l'art.

Bibliographie

BERA M. (2017), « Big Data et anonymisation », *Le CNAM Mag'*, N°8, Septembre.

EXECUTIVE OFFICE OF THE PRESIDENT (2014), *Big Data: Seizing opportunities, preserving Values*, The white House, Washington, mai. https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

EXECUTIVE OFFICE OF THE PRESIDENT & PRESIDENT'S COUNCIL OF ADVISORS ON SCIENCE & TECHNOLOGY (2014), *Big Data and Privacy: A Technological Perspective*, The White House, Washington, mai. https://bigdatawg.nist.gov/pdf/pcast_big_data_and_privacy_-_may_2014.pdf

GARNIER A. (2013), Livre blanc *Big Data et Réseaux Sociaux : Mythes & Réalités – La déclinaison pour les Réseaux Sociaux d'Entreprise*.

HAMEL M.-P. & MARGUERIT D. (2013), « Analyse des Big Data. Quels usages ? Quels défis ? », Commissariat général à la stratégie et à la prospective, *La note d'analyse*, n°08, 12 novembre. https://www.strategie.gouv.fr/publications/analyse-big-data-usages-defis

HENKE N., BUGHIN J., CHUI M. *et al.* (2016), *The Age of Analytics: competing in a Data-Driven world*, McKinsey Global Institute, décembre. https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world

HUOT C. *et al.* (2015), « Données massives : Big Data, Impact et attentes pour la normalisation », *Livre blanc*, Afnor Normalisation (Mars).

ISO / IEC JTC 1 (2014), "big_data_report-jtc1.pdf", *Information Technology*.

MANYIKA J., CHUI M., BROWN B. *et al.* (2011), *Big Data: the next frontier for innovation, competition, and productivity*, McKinsey Global Institute, Mai. https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

SWEENEY L. (2002), "Achieving k-anonymity privacy protection using generalization and suppression. " International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 571-588.