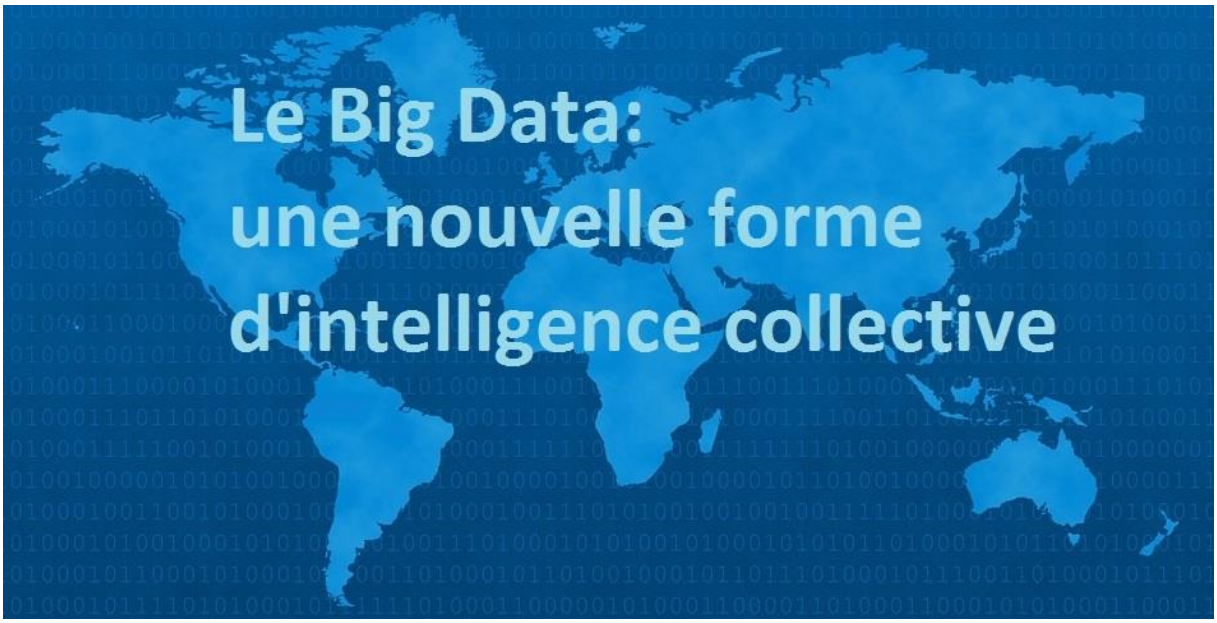

BIG DATA : UNE NOUVELLE FORME D'INTELLIGENCE COLLECTIVE

Mini-série de Billets #4 : Big Data – Sur les épaules de géants

Sur les épaules de géants…. Au fil de l'écriture de cette mini-série de Billets, et de mes plongées dans les eaux de l'histoire des mathématiques et de l'informatique, j'ai été frappée par le fait que les évolutions technologiques qui s'installent dans nos sociétés et les transforment, s'enracinent en d'autres temps… Sir Isaac Newton a prononcé cette phrase magnifique d'humilité: *« If I have seen further it is by standing on ye sholders of giants »*[i]. Notre voyage en Terra Data s'accomplit sur les épaules de géants… Nous découvrirons ce que le Big Data doit à d'autres géants, à l'un d'eux en particulier, et suisse de surcroît : Leonhard Euler, qui en résolvant le casse-tête des *7 ponts de Königsberg*[ii] a posé les prémisses d'une pierre angulaire du Big Data: la théorie des graphes.

Les thèmes abordés dans cette mini-série de Billets #4 seront les suivants :

- Variété, le 3e V du Big Data
- L'apport des mathématique dans le développement des moteurs de recherche et la genèse du Big Data – la théorie des graphes et la topologie
- Mais au fait, que cherche-t-on dans toutes ces données ?
- Tout cela est vraiment Big ! Big Data ?

Billet #4A : Variété, le 3e V du Big Data

Tout grand volume de données ne constitue pas en soi des Big Data, c'est-à-dire des structures informatiques capables de gérer des immenses volumes de données variées. Il existe trois dimensions que l'on retrouve dans la définition habituelle[iii] du Big Data : le *Volume*[iv], la *Vélocité*[v] et la *Variété* des données. 3V qui aident à dessiner les contours de l'expression *Big Data*, et de son contenu. Un des objectifs poursuivis à travers ces Billets.

Penchons-nous ici sur le V de *Variété*.

Des données d'un type nouveau

Avec le développement d'Internet, une masse de données d'un type nouveau a été générée.

Les données constituant le Big Data ne ressemblent en rien à un tableau Excel géant, avec ses données bien ordonnées à l'instar de ces véhicules parfaitement alignés….

Bien au contraire, composé de données bruyantes, *« il faut plutôt se figurer le Big Data comme un torrent de montagne, dont chaque goutte est un chiffre, ou encore comme une photo ou une suite de photos »*[vi]. Chiffres, textes, photos, musiques, vidéos nourrissent ce torrent. Autant de données de formats nouveaux, non structurées.

Traiter ces données dépasse le cadre des statistiques traditionnelles. Or, la nécessité d'améliorer l'efficacité des moteurs de recherches requiert de trouver de nouvelles façons de stocker et d'effectuer le traitement des données. Les solutions appliquées par le passé ne sont pas envisageables précisément en raison du volume conséquent de ces données, de leurs formats variés, mais également de leur caractère dynamique rendant impossible leur ordonnancement dans un tableau Excel[vii], dans une base de données traditionnelles[viii].

Nous avons vu dans la mini-série de Billets #3 quelles solutions ingénieuses ont été élaborées : architectures distribuées de type cluster d'ordinateurs, stockage distribué, répliqué et traitement parallèle des données. Fruit du constat que traiter ces gros volumes de données de formats variés exige de trouver des solutions innovantes, *Hadoop* a mis au point un environnement de données non structurées, *Hadoop Distributed System File* (HDSF), construit à partir d'un modèle massivement parallèle. Puis l'éléphant-*Hadoop* a grandi pour devenir aujourd'hui un véritable écosystème omniprésent[ix].

Des données de formats variés

Désormais les données proviennent de sources toujours plus variées et abondantes. Mentionnons seulement l'*Internet des objets* (IdO[x]) et les réseaux sociaux qui se développent tentaculairement.

Variété des données… Données complexes versus données relationnelles traditionnelles Données publiques ou privées, provenant d'acteurs divers, de formats hétérogène : texte, chiffres, sons, images, vidéos, données dynamiques (changeant constamment) ou statiques (qui ne changent pas[xi]), complètes ou incomplètes, qui se présentent sous un format différent de celui désiré (en raison de l'unité de taille, de poids, etc. d'origine, par exemple).

Ces données sont brutes et ainsi que*« leur nom l'indique, ne sont pas structurées ni affinées de quelque manière que ce soit. Il se peut que certaines données soient manquantes, ou endommagées, ou simplement, qu'elles ne conviennent pas »* [xii].

Les données sont fréquemment qualifiées en référence à leur structure : données structurées, semi-structurées ou non structurées. L'élément attributif à l'une ou l'autre carégorie étant la *base de données*, soit *« une collection d'informations organisées afin d'être facilement consultables, gérables et mises à jour »*[xiii]. Et Wikipédia de préciser : « *« conteneur » stockant des données telles que des chiffres, des dates ou des mots, pouvant être retraités par des moyens informatiques pour produire une information »*[xiv]. A titre d'exemple, on pensera aux bases de données des entreprises. Des lignes et des colonnes, des cellules en guise d'intersection.

I) Données structurées

D*onnées structurées* s'entend des données *« stockées, traitées et manipulées dans un système de gestion de base de données relationnelle traditionnelle (SGBDR) »*[xv]. Elles sont *« organisées dans un format facilement utilisable par une base de données ou une autre technologie. (....) Les données correspondent à des champs déterminés. Cela signifie que ceux qui utilisent des données peuvent prévoir de disposer d'éléments d'information de longueur fixe et de modèles cohérents, constants afin de traiter cette information. Dans le passé, les données structurées étaient la norme, et une grande partie de la structuration des données a été réalisée par des humains »*[xvi].

Les données structurées, chiffrées et prévisibles renvoient, par exemple, à l'âge du client, au code postal de son domicile, au chiffre d'affaires de l'entreprise, à l'historique des ventes. Si celles-ci sont utiles et facilement analysables, un nouveau type de données - les données non structurées - lesquelles permettent notamment d'accéder aux sentiments des consommateurs, les supplantent peu à peu[xvii].

II) Données non structurées

Les *données non-structurées* se réfèrent à des données *« qui sont couramment générées à partir d'activités humaines et qui ne s'intègrent pas dans un format de base de données structurées »*[xviii]. Ainsi, à la différence des données stockées dans des tableurs ou des bases de données, *« les données non structurées sont des données représentées ou stockées sans*

format prédéfini»[xix]. Elles sont constituées de *«toutes les données qui n'ont pas de structure reconnaissable. Elles sont non organisées et brutes et peuvent être non textuelles ou textuelles. (…). Elles suivent un format moins ordonné que des éléments comme les pages de tableur, les tables de base de données ou d'autres ensembles de données linéaires ou ordonnés. En fait, le terme "ensemble de données" est utile parce qu'il est associé à des données qui se présentent sous forme de tableaux accessibles, sans contenu supplémentaire, et qui sont liées ou étiquetées dans une structure spécifique»*[xx].

Devenu un élément incontournable de notre vie quotidienne*, « le courriel constitue une belle illustration de données textuelles non structurées. Il comprend l'heure, la date, les détails et l'objet du destinataire et de l'expéditeur, mais un corps d'email reste non structuré. Les données non structurées peuvent également être identifiées comme des données mal structurées, dans lesquelles les sources de données comprennent une structure, mais toutes les données d'un ensemble de données ne suivent pas la même structure»*[xxi]*.* Les exemples de données textuelles non structurées sont multiples: *« les documents Word, les présentations PowerPoint, les messages instantanés, les logiciels collaboratifs, les documents, les livres, les médias sociaux et dossiers médicaux. Les données non textuelles non structurées sont généralement créées dans les médias, tels que les fichiers audio MP3, les images JPEG et les fichiers vidéo Flash »*[xxii]. Les médias sociaux constituent une source intarissable de données non structurées.

III) Données semi-structurées

Il existe enfin des *données semi-structurées*, soit *« des données qui ne s'intègrent pas dans un système de base de données structuré, mais qui sont néanmoins structurées par des étiquettes (mots-clés) utiles pour créer un format ou ordonner une hiérarchie dans les données »*[xxiii]. Ni brutes, ni saisies dans un système de base de données conventionnel, il s'agit de données structurées, mais qui ne sont pas organisées selon un modèle rationnel[xxiv].

Ces définitions apportent un éclairage supplémentaire sur ce que j'avais écrit en page 7 du Billet #3D - - *Hadoop – une évolution: un écosystème, le Big Data & la résolution des conflits*, à propos des approches de stockage et de traitement des données centralisée, toujours plus inadaptée, systèmes insuffisants en terme de volume, mais également sous l'angle de la variété des données actuellement produites qui ne peuvent pour nombres d'entre elles être ni stockées ni traitées dans des bases de données relationnelles traditionnelles, principe prévalant avant le Big Data[xxv]. Maintenant, on comprend pourquoi !

Des données exploitables

Variété des données... *« A partir du moment où les données sont brutes, on ne peut pas être sûr de ce que l'on obtiendra et de ce que l'on pourra faire »*[xxvi]. *Les données*, pétrole du 21e siècle, qui comme l'or noir réalise son potentiel au-delà du raffinage, *faire parler les données* - données à faible densité informative[xxvii] - dont la valeur apparaît de la variété et du volume.

Exploiter ces données qui se présentent sous des formats variés, proviennent de supports divers, sont produites par de multiples acteurs[xxviii] nécessite toutefois de mettre un peu d'ordre dans cette déferlante. Avancer dans cette direction, c'est se souvenir que le Big Data repose sur les ordinateurs. Lesquels exécutent des algorithmes. Ainsi, la plupart du temps, il convient en premier lieu de structurer les données, les structurer précisément pour les rendre exploitables ; cela parce que la structure constitue un élément fondamental du fonctionnement d'un algorithme[xxix] et d'un ordinateur[xxx], structurer les données afin d'obtenir une solution, du fait que l'ordinateur est simple, non subtile : *« un ordinateur conçoit les données de façon structurée, simple, inflexible, et certainement pas créative »*[xxxi].

Conçu pour traiter des nombres uniquement, il a des données une vision simple, pour lui tout est nombre, il ne comprend ni les données, ni ce qu'elles impliquent : *« Pour un ordinateur, la lettre A n'est pas autre chose que le nombre 65. En fait, ce n'est même pas vraiment le nombre 65. Pour l'ordinateur, c'est une série d'impulsions électriques qui équivaut à la valeur en chiffres binaires 0100 0001 »*[xxxii]. Un ordinateur ne raisonne qu'en binaire. Nous sommes bien loin de la complexité de l'esprit humain. Cet exemple démontre que réussir le virage numérique implique de développer la pensée computationnelle (*computational thinking*), d'être capable de raisonner sur un problème de sorte qu'un ordinateur puisse le résoudre en exécutant des algorithmes.

Et parce que résoudre un problème est précisément ce que l'on attend d'un algorithme - *« succession d'étapes destinées à résoudre un problème. (…) séquence représentant une méthode unique de résolution d'un problème par la production d'une solution »*[xxxiii] - structurer des données signifie *« les organiser d'une certaine manière de telle sorte qu'elles aient les mêmes caractéristiques, la même apparence et les mêmes composantes (…). L'organisation constitue une partie importante de la structuration des données. Il ne s'agit pas du tout de modifier les données, seulement de les rendre plus exploitables »*[xxxiv].

Il existe différentes structures organisationnelles. On parle, notamment, de piles, de files, de listes, de dictionnaires, d'arborescences et de graphes. Avant de nous pencher au Billet #4B sur ces deux derniers éléments - arborescences et graphes - j'invite toute personne désireuse d'explications plus détaillées sur ces diverses structures organisationnelles à consulter l'ouvrage de John Paul Mueller et Luca Massaron, *Les algorithmes pour les nuls,* thèmes qui excèdent le sujet de cette série de Billets[xxxv].

Mais au fait, que cherche-t-on dans toutes ces données ?

J'ai placé cette question au cœur de cette mini-série de Billet #4, comme un point de départ de plusieurs interrogations indispensables, capitales, essentielles. Existentielles ?

Du sens, on cherche du sens. De l'information, de la connaissance ; cet objectif demeure inchangé. Une base de données traditionnelle permet de stocker, d'organiser des données d'un certain format et d'en extraire de l'information, *de les faire parler* grâce aux statistiques traditionnelles. Le but demeure, les moyens, les méthodes quant à eux évoluent, émergent. L'apparition de nouveaux formats de données

variées et volumineuses a posé de nouveaux problèmes, impliquant de changer d'approches pour les stocker et les traiter. Innover pour *faire parler les données*.

On perçoit aisément l'opportunité que constitue cette abondante variété de données volumineuses. Elle ouvre l'accès à de nouvelles informations en tout genre, à rendre visible des tendances inédites, pour qui sait faire parler les données, et cela de plus en plus vite. Les faire parler, vite, en temps réel y compris, car nombre d'entre elles se périment à peine produites[xxxvi].

S'agit-il là de l'émergence d'une nouvelle forme d'intelligence collective, distillée de l'intarissable production de données liée à la diversification croissante des acteurs engagés, des formats et des supports de données ?

Combien de pépites de connaissances ces données, miettes de nos vies laissées derrière nous, dissimulent-elles ?

Nouvelle approche, *« leur variété permet d'utiliser les grandes données pour faire des découvertes scientifiques sans suivre la méthode scientifique »*[xxxvii]. Nous reviendrons sur cet élément méthodologique, épistémologique, essentiel plus avant dans cette mini-série de Billets #4.

Du sens, des informations, des connaissances. Mais pour quels usages…. ? *Telle est la question*, aurait spécifié Hamlet.

Prémices à cette exploration, structurer les données, les organiser, faire un peu d'ordre, pour les rendre exploitables, faute de quoi, elles ne parleront pas, resteront à jamais muettes, gardiennes de leurs secrets, de leurs trésors. Dans ce contexte s'inscrit le Billet #4B. Quand bien même elle ne se veut pas exhaustive, l'étude de l'évolution des moteurs de recherche au Big Data – fil rouge de cette tentative de définition du Big Data - serait amputée d'un élément fondamental si je passais sous silence un apport essentiel des mathématiques - la théorie des graphes et la topologie - à la déferlante de données, à son organisation et à son exploitation, inscrits dans et modifiant notre quotidien.

A bientôt pour le Billet #4B !

Anne-Sylvie Weinmann

Références :

[i] En anglais moderne : *« If I have seen further it is by standing on the shoulders of giants »* (*Isaac Newton*, https://en.wikiquote.org/wiki/Isaac_Newton).

[ii] Ce sujet sera développé dans le Billet #4B à venir ; *Problème des sept ponts de Königsberg*, https://fr.wikipedia.org/wiki/Probl%C3%A8me_des_sept_ponts_de_K%C3%B6nigsberg

[iii] Définition inspirée d'un rapport du META Group (devenu Gartner) du Big Data en 3V. Au minimum 3V : Volume, Vélocité, Variété. On rencontre parfois des V additionnels pour *Valeur* et *Véracité* (*Big Data*, https://fr.wikipedia.org/wiki/Big_data; MUELLER John Paul, MASSARON Luca, *Les algorithmes pour les nuls*, Paris, Editions First, 2017, pp. 266-267 (https://www.eyrolles.com/Informatique/Livre/les-algorithmes-pour-les-nuls-9782412025901); DELORT Pierre, *Le Big Data*, Paris, Presses universitaires de France, 2015, p. 49).

[iv] Sur le Volume en général, voir le Billet #2 – *Données : des tablettes mésopotamiennes à nos tablettes numériques*. Je reviendrai ultérieurement au Billet #4D sur ce sujet, car tout grand volume de données n'est pas *Big* au sens de Big Data.

[v] A propos de la Vélocité en général, voir le Billet #3D - *Hadoop – une évolution: un écosystème, le Big Data & la résolution des conflits*, pp. 14ss (p. 15 en particulier). Je reviendrai ultérieurement au Billet #4D sur ce sujet, car toute grande Vélocité de données n'est pas *Big* au sens de Big Data.

[vi] BABINET Gilles, *Big Data : penser l'homme et le monde autrement*, Paris, Le Passeur, 2015, p. 32 (http://www.eyrolles.com/Informatique/Livre/big-data-penser-l-homme-et-le-monde-autrement-9782368904923).

[vii] Un tableau Excel comporte 1'048'576 lignes et 16'384 colonnes (https://support.office.com/fr-fr/article/Sp%C3%A9cifications-et-limites-relatives-%C3%A0-Excel-1672b34d-7043-467e-8e27-269d656771c3#ID0EBACAAA=Excel%C2%A02016-2013).

[viii] Billet #3A - *Des moteurs de recherche au Big Data. Et la résolution des conflits ?*, p. 3.

[ix] Voir la mini-série de Billets #3 ; CHOKOGOUE Juvénal, *Hadoop : devenez opérationnel dans le monde du Big Data*, St-Herblain, ENI, 2017 (https://m.editions-eni.fr/livre/hadoop-devenez-operationnel-dans-le-monde-du-big-data-9782409007613#); et le nouvel ouvrage de CHOKOGUE Juvénal, *Maîtrisez l'utilisation des technologies Hadoop*, Paris, Eyrolles, 2018 (https://www.eyrolles.com/Informatique/Livre/maitrisez-l-utilisation-des-technologies-hadoop-9782212674781).

[x] Voir le Billet #3D - *Hadoop – une évolution: un écosystème, le Big Data & la résolution des conflits*, pp. 14ss.

[xi] MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 21.

[xii] MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 133.

[xiii] L. Bastien, *Bases de données : qu'est-ce que c'est ? Définition et Présentation*, 07/05/2018 sur Le Big Data (https://www.lebigdata.fr/base-de-donnees).

[xiv] *Base de données relationnelle*, https://fr.wikipedia.org/wiki/Base_de_données_relationnelle; sous la même référence, l'encyclopédie en ligne ajoute : *« Une base de données (database en anglais), permet de stocker et de retrouver l'intégralité de données brutes ou d'informations en rapport avec un thème ou une activité ; celles-ci peuvent être de natures différentes et plus ou moins reliées entre elles. Dans la très grande majorité des cas, ces informations sont très structurées, et la base est localisée dans un même lieu et sur un même support. Ce dernier est généralement informatisé »*.

[xv] Traduction de : PIERSON Lillian, *Data Sciences for dummies*, Hoboken (USA), John Wiley & Sons, 2e éd., 2017, p. 8 (http://eu.wiley.com/WileyCDA/WileyTitle/productCd-1119327636.html). L'acronyme *SGBDR* signifie : système de gestion de base de données relationnelle. Selon Wikipédia, *« en informatique, une base de données relationnelle est une base de données où l'information est organisée dans des tableaux à deux dimensions appelés des relations ou tables, selon le modèle introduit par Edgar F. Codd en 1970. Selon ce modèle relationnel, une base de données consiste en une ou plusieurs relations. Les lignes de ces relations sont appelées des nuplets ou enregistrements. Les colonnes sont appelées des attributs. Les logiciels qui permettent de créer, utiliser et maintenir des bases de données relationnelles sont des systèmes de gestion de base de données relationnels. Pratiquement tous les systèmes relationnels utilisent le langage SQL pour interroger les bases de données. Ce langage permet de demander des opérations d'algèbre relationnelle telles que l'intersection, la sélection et la jointure »* (*Base de données relationnelle*, https://fr.wikipedia.org/wiki/Base_de_données_relationnelle).

[xvi] Traduction de : *Structured data*, https://www.techopedia.com/definition/30363/structured-data

[xvii] AURINE Guillaume, *Big Data : à la conquête des données non structurées*, 22/08/2016 sur SalesForce Blog (https://www.salesforce.com/fr/blog/2016/08/big-data-a-la-conquete-des-donnees-non-structurees.html).

[xviii] Traduction de : PIERSON Lillian, *op. cit.*, p. 8.

[xix] *Données non structurées*, https://fr.wikipedia.org/wiki/Donn%C3%A9es_non_structur%C3%A9es

[xx] Traduction de : *Unstructured data*, https://www.techopedia.com/definition/13865/unstructured-data

[xxi] Traduction de : *Unstructured data*, https://www.techopedia.com/definition/13865/unstructured-data

[xxii] Traduction de : *Unstructured data*, https://www.techopedia.com/definition/13865/unstructured-data

[xxiii] Traduction de : PIERSON Lillian, *op. cit.*, p. 8 ; *Semi-structured data*, https://en.wikipedia.org/wiki/Semi-structured_data

[xxiv] *Semi-structured data*, https://www.techopedia.com/definition/28802/semi-structured-data qui propose les exemples suivants : fichiers BibTex (*BibTeX*, https://fr.wikipedia.org/wiki/BibTeX) ou un document Standard Generalized Markup Language (SGML) (*Standard Generalized Markup Language*, https://fr.wikipedia.org/wiki/Standard_Generalized_Markup_Language).

[xxv] *« Avant le Big data, les données automatisées étaient traitées, souvent au sein d'un ordinateur central, à l'aide d'un SGBDR. Aujourd'hui, le volume de données générées dépasse largement la capacité de stockage et de traitement de tout serveur et de tout SGBDR. Dès lors, les approches de traitement et de stockage centralisées qui ont prévalu jusqu'ici sont simplement inenvisageables »* (CHOKOGOUE Juvénal, *op. cit*., 2017, p. 299). Pour en savoir plus sur les bases de données en général, cf. L. Bastien, *Bases de données : qu'est-ce que c'est ? Définition et Présentation*, 07/05/2018 sur Le Big Data (https://www.lebigdata.fr/base-de-donnees).

[xxvi] MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 133.

[xxvii] DELORT Pierre, *op. cit*, pp. 44 et 46.

[xxviii] AURINE Guillaume, *Big Data : à la conquête des données non structurées*, 22/08/2016 sur SalesForce Blog (https://www.salesforce.com/fr/blog/2016/08/big-data-a-la-conquete-des-donnees-non-structurees.html).

[xxix] MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 134.

[xxx] MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 26.

[xxxi] MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 24.

[xxxii] MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 25.

[xxxiii] MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 11.

[xxxiv] MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 133.

[xxxv] MUELLER John Paul, MASSARON Luca, *op. cit.*, pp. 133ss.

[xxxvi] Cf. Billet #3D - *Hadoop – une évolution: un écosystème, le Big Data & la résolution des conflits*, pp. 14 ss.

[xxxvii] MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 267.