Sûreté de Fonctionnement des Big Data. Opportunités and Challenges

Fatima Ezzahra Mdarbi, Nadia Afifi, Imane Hilal

Sûreté de Fonctionnement des Big Data

Opportunités & Challenges

Fatima EzzahraMdarbi[1], Nadia AFIFI[2], Imane HILAL[1,3]
[1]Laboratoire RITM, EST, CED Sciences de l'Ingénieur ENSEM, Université Hassan II, Casablanca, Maroc
[2]Laboratoire RITM, EST, Département Génie Informatique, Université Hassan II, Casablanca, Maroc
[3]Laboratiore Lyrica, ESI, Rabat, Maroc
fati.marbi@gmail.com, Nafifi@est-uh2c.ac.ma, Ihilal@esi.com,

RESUME : Les Big Data représentent un ensemble de données très volumineux, dont son analyse dépasse les capacités des systèmes de gestion de base de données classiques. Ils ont toujours été lié au besoin de grande capacité de calcul et de stockage des flux de données.
La sûreté de fonctionnement (SdF) des données est l'une des préoccupations majeures des organisations. Elle traduit la confiance qu'on peut accorder à un système. De nos jours, les entreprises trouvent un intérêt majeur au Big data, mais le problème de la SdF reste le frein principal.
Dans cet article on présente les différents travaux ayant déjà traité les aspects de la SdF des Big Data. Cette étude permet de mettre en exergue de nouvelles opportunités dans ce domaine ainsi que les différents challenges.

Mots clés : Sûreté de Fonctionnement, Big Data, Confidentialité, Disponibilité, Fiabilité, Sécurité, Intégrité, Maintenabilité, 5V.

1 INTRODUCTION

L'explosion quantitative des données numériques a obligé les chercheurs à trouver de nouvelles manières de voir et d'analyser le monde. Il s'agit de découvrir de nouveaux ordres de grandeur concernant la capture, la recherche, le partage, le stockage, l'analyse et la présentation des données ce qui a induit au Big Data.
Le terme Big Data fait référence à l'augmentation du volume des données qui sont difficiles à stocker, traiter et analyser avec des technologies traditionnelles.
Les utilisateurs des Big Data sont confrontés aux (i) problèmes de la scalabilité des systèmes, (ii) du temps de réponse aux requêtes des utilisateurs, (iii) de la sécurité des transactions, (vi) la fiabilité et disponibilité des résultats des traitements. Ces problèmes représentent certains aspects de la SdF qui doivent être pris en charge en amont par les plateformes Big Data. Ainsi pour cerner le contexte général des aspects de la SdF des Big Data, nous proposons dans cet article un état de l'art des travaux ayant déjà traité ce sujet.

Le présent article sera organisé comme suit : Dans la section 2nous présentons les principales caractéristiques des Big Data. La section 3 traite la SdF et ses attributs. Dans la section 4 les travaux qui ont été mené dans le contexte de la SdF des Big Data seront mis en exergue. Alors que la section 5 présente une analyse et une synthèse des différents travaux étudiés. Enfin nous terminons par une conclusion et quelques perspectives.

2 BIG DATA

Les Big Data est un terme utilisé pour décrire les données dans le réseau lorsqu'elles dépassent les capacités des systèmes traditionnelles[1]. Les Big Data font référence à l'augmentation du volume des données qui sont difficiles à stocker, traiter et analyser avec des technologies traditionnelles[2].Les caractéristiques intrinsèques des Big Data connues par les 5V sont : (i) Volume; (ii) Vitesse ; (iii) Variété ; (iv) Valeur ; (v) Véracité[3] comme le montre la figure 1[4].

Figure 1. 5V de Big data

Le Volume : représente la quantité de données générée. Cette dernière est en pleine expansion et suit une loi quasi exponentielle. Ceci est dû au commerce électronique et aux réseaux sociaux qui contribuent à cette profusion de données.

La Vitesse : représente la fréquence à laquelle les données sont générées, le traitement des données doit se faire souvent en temps réel.

La Variété : caractérise les données qui peuvent prendre des formes très variées et très hétérogènes (voix, données faciales, données transactionnelles, web analytique, textes, images, etc.)

La Valeur : représente l'utilité des données. La valeur des données vient de l'analyse et de l'application des résultats aux besoins spécifiques.

La Véracité: conditionne quant à elle directement la pertinence de la donnée[5].Elle est notamment menacée par les comportements déclaratifs et par les diversités des points de collecte. La multiplication des formats de données et l'activité des robots et faux profils innombrables sévissant sur Internet contribuent à cette menace.

3 SURETE DE FONCTIONNEMENT

La SdF consiste à évaluer les risques potentiels, prévoir l'occurrence des défaillances et tenter de minimiser les conséquences des situations catastrophiques lorsqu'elles se présentent[6]. Selon Zwingelstein[7], la SdF est l'aptitude d'une entité à réaliser une ou plusieurs fonctions requises dans des conditions données. Elle peut être appliquée au niveau d'un processus, d'un système, d'un composant, suivant la profondeur de l'analyse.

Les concepts fondamentaux de la SdF sont classés en trois groupes. La figure 2 illustre l'arbre de la SdF selon Laprie [8].

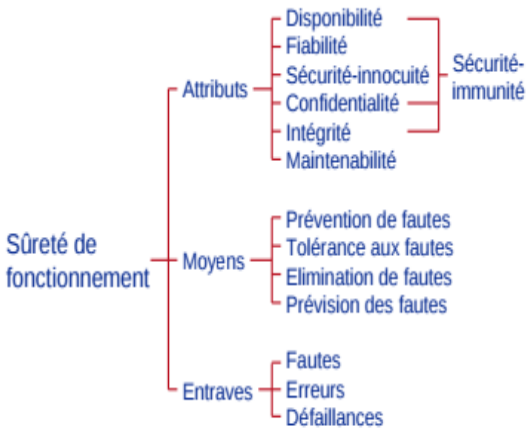

Figure 2. Arbre de la SdF

La SdF manipule un certain nombre de concepts ; Elle peut être aperçue comme étant composée des trois éléments suivants :(i) les attributs, (ii) les entraves, et (iii) les moyens. Les entraves sont liées aux circonstances indésirables mais non inattendues. Les moyens correspondent aux méthodes et techniques permettant de garantir l'aptitude du système à délivrer un service conforme à l'accomplissement de sa fonction. Ils permettent également de faire confiance à cette aptitude. Quant aux attributs, ils permettent d'exprimer les propriétés attendues du système et d'apprécier la qualité du service délivré[9]. Ces attributs sont :

Disponibilité : représente l'aptitude d'une entité à être en état d'accomplir une fonction requise dans des conditions données et à un instant donné.

Fiabilité : représente l'aptitude d'un dispositif à accomplir une fonction requise dans des conditions données pendant une durée donnée.

Sécurité innocuité : c'est l'aptitude d'une entité à éviter de faire apparaitre, dans des conditions données, des événements critiques

Intégrité : représente la non-occurrence d'altérations inappropriées du système.

Maintenabilité : c'est la faculté d'une entité à être maintenue ou rétablie

Dans cet article, nous n'allons nous intéresser qu'aux attributs de la SdF à savoir : la confidentialité, la disponibilité, la fiabilité, la sécurité, l'intégrité, et la maintenabilité. Nous discuterons également des travaux scientifiques ayant fait des propositions dans ce sens.

4 INTERFERENCE DE SDF & BIG DATA

Vue l'évolution importante des Big Data, la SdF représente de nos jours un intérêt majeur pour l'utilisateur. C'est pourquoi différent auteurs ont discuté de la relation Big Data et SdF. Dans ce sens, Wu et al ont proposé un modèle mathématique pour caractériser la fiabilité des Big Data. Ce modèle utilise l'entité problème et l'entité donnée pour représenter le problème et les données correspondants dans le monde réel[10].

Selon [11] un nouveau schéma de signatures généralisé basé sur l'ID en utilisant les techniques de Waters [12] et de CHK [13], a été proposé pour améliorer la confidentialité et l'authenticité des Big Data. Le schéma proposé peut fonctionner comme un schéma de cryptage, un schéma de signature ou un schéma signcryption selon le besoin.

Machanavajjhala et Reiter ont défini et mesuré les risques de confidentialité, ainsi que les méthodes de protection des données de diffusion publique[14].

Les auteurs dans [15] soulignent les dix principaux défis en matière de sécurité et de confidentialité des Big Data. La mise en évidence des défis motivera l'accent mis sur la consolidation des infrastructures Big Data.

Meeker et Hong présentent un état de l'art sur certaines applications traitant la fiabilité des données. Ils explorent la possibilité d'utiliser la fiabilité des données pour fournir des méthodes statistiques plus solides afin d'exploiter et prédire les performances des systèmes sur le terrain [16].

Les travaux de Tankard, développent une approche au niveau de la sécurité, qui consiste à contrôler les Big Data. Les contrôles doivent être au niveau des données et au centre de données, afin de fournir une ligne de défense plus efficace. En ce qui concerne les contrôles d'accès, ils doivent être suffisamment granulaires pour que seules les personnes autorisées à accéder aux données puissent le faire, afin d'éviter que des

informations sensibles ne soient compromises. Les contrôles doivent également être définis selon le principe des privilèges minimaux, en particulier pour les utilisateurs disposant de droits d'accès plus importants, tels que les administrateurs. Une telle approche est très efficace dans tout environnement multi-silo où toute forme de données électroniques est stockée[17].

Kumar et al. Proposent une approche pour exploiter le potentiel des Big Data et faciliter la construction et la maintenance des systèmes axés sur l'analyse des données volumineuses [18].

Quant à Patil et Seshadri , eux ils présentent les problèmes de pointe en matière de sécurité et de confidentialité des Big Data appliqués dans le secteur de la santé[19].

En ce qui concerne la sécurité Lafuente propose les techniques suivantes :

1- Cryptage des données : Le cryptage est bien sûr la solution principale pour garantir une protection des données lors du stockage.
2- Contrôle d'accès et surveillance : Des mécanismes de contrôle d'accès adéquats seront essentiels pour protéger les données. Le contrôle d'accès est traditionnellement fourni.
3- Approches de politiques et de conformité : Utilisation des kits de conformité conçus pour fonctionner dans un environnement Big Data.
4- Gouvernance et cadres : Si aucun cadre de gouvernance n'est pas appliqué aux Big Data, les données collectées pourraient être trompeuses et entraîner des coûts inattendus. [20].

Dans leurs traveaux Matturdi et al. élaborent un état d'art sur les énormes avantages et défis de la sécurité et de la confidentialité des Big Data. Ensuite, ils présentent quelques méthodes et techniques possibles pour assurer la sécurité et la confidentialité des Big Data [21].

Dans [22], une méthode de modélisation des systèmes d'entrepôts qui garantit la SdF des données a été proposé. Cette méthode est orientée aspect et basée sur l'approche MDA (Model Driven Architecture).

Moghadam et Colomo-Palacios, proposent une cartographie systématique sur les aspects d'interférences entre les Big Data et la gouvernance de la sécurité de l'information. Ils illustrent les défis et les lacunes concernant le sujet et clarifient ces défis à l'aide : (i) d'une classification des environnements dans lesquels ils se déroulent ;(ii) des spectres de risques de sécurité concerné par ces Big Data ; (iii)et des mesures de gouvernance de la sécurité qu'ils prennent pour les atténuer. Ceci est réalisé grâce à des solutions sous forme de framework, modèle, logiciel ou autre outil. Les résultats devraient être utiles aux professionnels de la sécurité informatique et aux professionnels des systèmes d'information dans leur ensemble [23].

Une approche pour sécuriser le modèle de menace pour le grand cycle de vie des données de santé a été présentée par Khaloufi et al. Ils ont aussi mis l'accent sur la description des techniques récemment proposées concernant (i) l'authentification, (ii) le cryptage,(iii) l'anonymisation, (iv) le contrôle d'accès, (v) et la confidentialité. Ainsi , ils ont proposé des politiques et des mécanismes globaux visant à résoudre les différents problèmes de sécurité des Big Data dans le domaine de la santé [24].

Wang et al. Disent que le cryptage à base d'attributs est recommandé par la Cloud Security Alliance (CSA) comme l'un des outils de cryptographie possibles pour le contrôle d'accès dans les applications Big Data. Dans ABE, le fichier partagé ne peut être chiffré avec la stratégie spécifique qu'une seule fois ; et il peut être déchiffré par tout destinataire dont les attributs sont satisfaits. Lorsque ABE est déployé dans certains scénarios de réseau ouvert, il est inévitablement attaqué par des attaques par canaux parallèles, du fait que les Big Data proviennent de différents points de terminaison. Aussi ils proposent des schémas Ciphertext-policy attribute based encryption (CP-ABE) et Key-policy attribute based encryption (KP-ABE) résilients aux fuites dans le modèle d'entrée auxiliaire amélioré, qui permet à l'attaquant d'obtenir plus d'informations sur les fuites concernant le caractère aléatoire du chiffrement ; et ce après avoir vu le texte chiffré d'interrogation. De plus, ils construisent un extracteur puissant amélioré à partir du théorème de Goldreich Levin [25] modifié pour la preuve de sécurité [26]. Karkouda et al., 2012, proposent une façon pour limiter ces risques protéger les entrepôts de données des différents risques et dangers qui sont nés avec le stockage dans le Cloud. Ceci à travers l'algorithme de partage de clés secret de Shamir [27]. Une étude comparative sur la SdF des Big Data implémentées dans les environnements Cloud [4].

5 ANALYSE

Nous avons présenté dans la partie précédente une étude de quelques travaux qui proposent des solutions pour résoudre les problèmes liés à la SdF des Big Data. Comme présenté dans le tableau 1, et d'après notre étude nous constatons que la plupart des travaux n'ont traité qu'un seul attribut de la SdF. On peut citer que dans les articles [17], [18], [19], [20], [21] des solutions pour améliorer la sécurité des Big Data ont été présentées. Alors que d'autres auteurs dans [9], [12] à titre d'exemple, ont proposé des approches pour assurer la fiabilité. La confidentialité quant à elle a été abordé dans les travaux [10] et [11]. Concernant la maintenabilité, une approche a été proposée dans l'article [15]. On trouve également des travaux ayant fait une combinaison de deux attributs comme proposé dans [13], [16] et [18] qui ont traité la confidentialité et la sécurité. Par contre selon nos recherches nous n'avons pas trouvé des travaux qui traitent

l'interférence Big Data avec les cinq attributs de la SdF à savoir : (i) Disponibilité, (ii) Confidentialité, (iii) fiabilité, (iv) Sécurité, (v) Intégrité, (vi) Maintenabilité.

On constate que, généralement les solutions proposées pour sécuriser les traitements des données sont basées essentiellement sur la cryptographie. Cette solution n'est pas toujours complète et ne permet pas de protéger les données de manière exclusive. En plus le mécanisme de cryptage et de décryptage des données peut être intensive, dans certains cas, sur les processeurs ce qui peut engendrer une grande sollicitation des ressources.

Tableau 1. Travaux Relatifs aux Big Data & SdF

Attribut / Article	Disponibilité	Confidentialité	Fiabilité	Sécurité	Intégrité	Maintenabilité
[10]			x			
[11]		x				
[14]		x				
[15]		x		x		
[16]			x			
[17]				x		
[18]						x
[19]		x		x		
[20]				x		
[21]		x		x		
[22]				x		
[23]				x		
[25]				x		
[27]	x	x				

La montée exponentielle des Big Data présente à la fois de grandes opportunités et de grands défis. Les opportunités incluent : (i) une accélération du traitement des données, (ii) de meilleurs moyens pour organiser l'information, (iii) des décisions commerciales mieux éclairées, (iv) une gestion plus efficace de la chaîne d'approvisionnement et de la répartition des ressources. Alors que les défis demeurent énormes : d'une part, les données se présentent sous différentes formes: texte, audio, vidéo, OCR (Reconnaissance Optique de Caractères), données de capteurs, etc. Et d'autre part, la prolifération d'algorithmes évolutifs pour obtenir des informations à partir de ces données peut s'avérer décourageante pour un utilisateur possédant un ensemble de données particulier.

6 CONCLUSION

De nos jours la SdF des Big Data représente un aspect important pour toutes les organisations. Vu l'importance cruciale des données et la croissante des menaces auxquelles elles sont confrontées, ce sujet suscite un intérêt croissant de la part de la communauté scientifique. D'un autre côté, les Big Data sont en train de se généraliser à mesure que l'informatique gagne en importance pour les organisations du monde entier. Afin d'avoir une meilleure visibilité des aspects de la SdF qui interférent avec les Big Data, nous avons présenté dans cet article une étude des différents travaux de recherche qui traitent ce sujet. Cette étude nous a amené à conclure que, dans la majorité des cas les chercheurs traitent un ou deux attributs de SdF (Sécurité/confidentialité) qui sont en relation avec les Big Data.

D'où l'idée de proposer des travaux qui tiennent en compte la combinaison de plusieurs attributs (Disponibilité, Confidentialité, fiabilité, Sécurité, Intégrité, Maintenabilité) de la SdF. Afin d'avoir des Big Data qui répondent aux exigence de la SdF.

Dans nos futurs travaux, nous prévoyons d'automatiser le processus de choix d'une plateformes Big Data dependable.

BIBLIOGRAPHIE

[1] R. Materese, « NIST Big Data Interoperability Framework: Volume 1, Definitions », *NIST*, 22-oct-2015. [En ligne]. Disponible sur: https://www.nist.gov/node/790381. [Consulté le: 16-nov-2016].

[2] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, et S. Ullah Khan, « The rise of "big data" on cloud computing: Review and open research issues », *Information Systems*, vol. 47, p. 98-115, janv. 2015.

[3] S. Yin et O. Kaynak, « Big Data for Modern Industry: Challenges and Trends [Point of View] », *Proceedings of the IEEE*, vol. 103, nº 2, p. 143-146, févr. 2015.

[4] F. E. Mdarbi, N. Afifi, et I. Hilal, « Comparative Study: Dependability of Big Data in the Cloud », in *Proceedings of the 2Nd International Conference on Big Data, Cloud and Applications*, New York, NY, USA, 2017, p. 19:1–19:6.

[5] Bruno Teboul et Thierry Berthier, « Valeur et Véracité de la donnée : enjeux pour l'entreprise et défis pour le Data Scientist. »

[6] C. Pagetti-ENSEEIHT, « Module de sûreté de fonctionnement », 2010.

[7] J. Gandibleux, « Contribution à l'évaluation de sûreté de fonctionnement des architectures de surveillance/diagnostic embarquées. Application au transport ferroviaire », phdthesis, Université de Valenciennes et du Hainaut-Cambresis, 2013.

[8] J.-C. Laprie, « Dependable Computing: Concepts, Limits, Challenges », *the 25th IEEE International Symposium on Fault-Tolerant Computing, Pasadena, California, USA*, 27-juin-1995.

[9] G. A. P. Castaneda, « Evaluation par simulation de la sûreté de fonctionnement de systèmes en contexte dynamique hybride », phdthesis, Institut National Polytechnique de Lorraine - INPL, 2009.

[10] X. Wu, X. Liu, et S. Dai, « The reliability of Big Data », in *2014 IEEE 7th Joint International Information Technology and Artificial Intelligence Conference*, 2014, p. 295-299.

[11] G. Wei, J. Shao, Y. Xiang, P. Zhu, et R. Lu, « Obtain confidentiality or/and authenticity in Big Data by ID-based generalized signcryption », *Information Sciences*, vol. 318, p. 111-122, oct. 2015.

[12] B. Waters, « Efficient Identity-Based Encryption Without Random Oracles », in *Advances in Cryptology – EUROCRYPT 2005*, 2005, p. 114-127.

[13] R. Canetti, S. Halevi, et J. Katz, « Chosen-Ciphertext Security from Identity-Based Encryption », in *Advances in Cryptology - EUROCRYPT 2004*, 2004, p. 207-222.

[14] A. Machanavajjhala et J. P. Reiter, « Big Privacy: Protecting Confidentiality in Big Data », *XRDS*, vol. 19, nº 1, p. 20–23, sept. 2012.

[15] « Expanded Top Ten Big Data Security and Privacy Challenges : Cloud Security Alliance ». [En ligne]. Disponible sur: https://cloudsecurityalliance.org/download/expanded-top-ten-big-data-security-and-privacy-challenges/. [Consulté le: 02-déc-2016].

[16] W. Q. Meeker et Y. Hong, « Reliability Meets Big Data: Opportunities and Challenges », *Quality Engineering*, vol. 26, nº 1, p. 102-116, janv. 2014.

[17] C. Tankard, « Big data security », *Network Security*, vol. 2012, nº 7, p. 5-8, juill. 2012.

[18] A. Kumar, F. Niu, et C. Ré, « Hazy: Making It Easier to Build and Maintain Big-data Analytics », *Queue*, vol. 11, nº 1, p. 30:30–30:46, janv. 2013.

[19] H. K. Patil et R. Seshadri, « Big Data Security and Privacy Issues in Healthcare », in *2014 IEEE International Congress on Big Data*, 2014, p. 762-765.

[20] G. Lafuente, « The big data security challenge », *Network Security*, vol. 2015, nº 1, p. 12-14, janv. 2015.

[21] B. Matturdi, X. Zhou, S. Li, et F. Lin, « Big Data security and privacy: A review », *China Communications*, vol. 11, nº 14, p. 135-145, Supplement 2014.

[22] I. Hilal, N. Afifi, H. Belhadaoui, H. R. Filali, et M. Ouzzif, « Aspect-Oriented Method for dependable data warehouse systems based on MDA approach », in *Multimedia Computing and Systems (ICMCS), 2014 International Conference on*, 2014, p. 1274–1278.

[23] R. S. Moghadam et R. Colomo-Palacios, « Information security governance in big data environments: A systematic mapping », *Procedia Computer Science*, vol. 138, p. 401-408, janv. 2018.

[24] H. Khaloufi, K. Abouelmehdi, A. Beni-hssane, et M. Saadi, « Security model for Big Healthcare Data Lifecycle », *Procedia Computer Science*, vol. 141, p. 294-301, janv. 2018.

[25] « gl.pdf ». .

[26] Z. Wang, C. Cao, N. Yang, et V. Chang, « ABE with improved auxiliary input for big data security », *Journal of Computer and System Sciences*, vol. 89, p. 41-50, nov. 2017.

[27] K. Karkouda, N. Harbi, J. Darmont, et G. Gavin, « Confidentialité et disponibilité des données entreposées dans les nuages », in *9ème atelier Fouille de données complexes (EGC-FDC 2012)*, Bordeaux, France, 2012.