

QUAND LES DONNEES MASSIVES ENVAHISSENT NOTRE QUOTIDIEN ET INFLUENCENT NOTRE COMPORTEMENT !

WHEN BIG DATA INVADE OUR DAILY AND INFLUENCE OUR BEHAVIOR !

MUKENGE MBUMBA Josich

Haute Ecole de Commerce de Kinshasa (République Démocratique du Congo)

Résumé : Les mutations actuelles sur le plan tant technologique que des besoins des utilisateurs ont poussé ces derniers à changer leur mode de production et de consommation des données. Nul n'ignore le fait que, sans données valorisées, il n'y aura pas d'informations de qualité et non plus de la connaissance pouvant éclairer les preneurs des décisions à pouvoir le faire en toute compétence. Et en parlant des données, il s'observe une production et une consommation sans limite dans le chef tant des entreprises que des utilisateurs. Raison pour laquelle, dans le cadre de cet article, nous avons commencé par expliquer les différentes dimensions que doivent, de nos jours, prendre les données produites afin de permettre leur consommation avec une valeur ajoutée dans les différentes circonstances. Nous avons par la suite, procédé par une enquête dans les milieux tant des professionnels informaticiens que des simples utilisateurs, afin de nous faire une idée sur ce constat. Et enfin nous avons proposé des pistes de solutions pour que les données quelle que soit leur finalité, puissent réellement contribuer à l'émergence de tous en leur procurant de la véritable compétence.

Mots clés : Données massives, Big Data, 5V, Hadoop, Spark, Echange données numériques, Sécurisation données.

Abstract : Current changes both technologically and in user needs have pushed them to change the way they produce and consume data. No one is unaware of the fact that, without valued data, there will be no quality information and no knowledge that can inform decision-makers to be able to do so with complete competence. And speaking of data, we observe unlimited production and consumption on the part of both companies and users. This is why, in the context of this article, we began by explaining the different dimensions that the data produced must take in order to allow their consumption with added value in different circumstances. We subsequently carried out a survey in the circles of both IT professionals and simple users, in order to get an idea of this observation. And finally we proposed possible solutions so that data, whatever their purpose, can truly contribute to the emergence of all by providing them with real skills.

Keywords : Massive data, Big Data, 5V, Hadoop, Spark, Digital data exchange, Data security

1. Introduction

De nos jours, les habitudes des utilisateurs des Nouvelles Technologies de l'Information et de la Communication subissent quotidiennement des mutations dues au fait de la production et de la consommation des masses importantes de données. Ces habitudes touchent toutes les couches de la société qui se voient, à cause ou grâce à la délocalisation des services et de la dématérialisation des documents, obligés de recourir à plusieurs technologies en vogue, pour augmenter leur productivité et ainsi répondre à tous leurs besoins.

L'importance accordée aux données dans cet article se justifie par le fait qu'elles constituent le socle sur lequel doit se baser toute personne physique ou morale pour avoir de la compétence, en passant par les informations de qualité et la connaissance. Ce qui, par ricochet, donnera aux organisations à la fois le pouvoir de faire face à la concurrence d'où qu'elle vienne en matière de production et de consommation des données, mais aussi et surtout la possibilité d'accroître leurs chiffres d'affaires dans un monde en perpétuelle mutation¹.

L'objectif poursuivi dans le cadre de cet article consiste à comprendre d'une part pourquoi la production et la consommation des données ont augmenté à une vitesse exponentielle, et d'autre part si les utilisateurs de ces dernières savent que les informations, la connaissance et la compétence sont les résultats issus des données bien traitées.

L'atteinte de cet objectif principal est matérialisée respectivement par les objectifs spécifiques ci-après :

- passer en revue les différentes dimensions des données massives et expliquer la procédure à suivre pour avoir respectivement les informations, la connaissance et la compétence ;
- présenter quelques technologies existantes pour exploiter les données massives ;
- étudier les besoins des utilisateurs sur base d'un échantillon représentatif choisi, et proposer des pistes de solutions.

2. Présentation des dimensions des données, des technologies, des méthodes d'évaluation et de la discussion

La démarche retenue dans cette approche méthodologique consiste respectivement à présenter les différentes dimensions des données, quelques technologies en vogue pour la manipulation des données massives, les instruments de collecte des données ; à déterminer la population et l'échantillon de l'étude ; et à présenter les résultats de l'enquête ainsi que la discussion.

2.1 Différentes dimensions des données

Les différentes dimensions des données consistent à répondre aux insuffisances en matière des données au niveau des Systèmes d'Information classiques². Il s'agit de cinq dimensions (Volume, Vitesse, Variété, Vérité et Valeur) reconnues sous l'abréviation 5V. Ci-après leur présentation.

¹ MUKENGE MBUMBA, J. (2024). QUID DE L'ECHANGE SECURISE DES DOCUMENTS DEMATERIALISES ?. *Revue internationale de responsabilité financière, d'économie, de gestion et d'audit (IJFAEMA)*, 6 (1), 131-141. <https://doi.org/10.5281/zenodo.10684117>

² Basés sur la manipulation des données structurées qui sont créées avec des systèmes de gestion des bases de données relationnelles.

2.1.1 Volume : augmentation sans cesse dans les Systèmes d'Information

La mondialisation oblige tout le monde à créer chaque jour des données, lesquelles sont conservées pour raison de gestion dans des historiques plus importantes. Mais aussi avec la dématérialisation des supports, ceci fait que les volumes des données ne cessent d'augmenter jusqu'à se chiffrer en Téraoctets. D'où l'incapacité pour les outils actuels de gestion de bases de données et de traitement des données à supporter une telle explosion de volume. Ils deviennent incapables de répondre aux nouveaux besoins dans des délais raisonnables et à des coûts raisonnables. Même s'il faut considérer les outils de Datamining³, ces derniers n'analysent de nombreuses données que sur un échantillon représentatif, alors que la gestion des données massives le fait sur l'intégralité des données.

Le temps étant de l'argent ; les volumes de données à traiter et à stocker de plus en plus importants dans les Systèmes d'Information peuvent rapidement devenir un handicap pour les entreprises (coûts exponentiels, incapacité d'analyser et d'accéder à toutes ces données dans un délai raisonnable, etc.). Ils peuvent par contre devenir une force par la mise en place des solutions capables de gérer et de valoriser ces volumes de données.

2.1.2 Vélocité : des données à collecter et à analyser en temps réel

En recourant aux capteurs et aux réseaux sociaux, les données sont produites continuellement, et doivent donc être traitées aussi en temps réel car elles s'intègrent pour certaines dans des processus sensibles au temps. Quant aux décideurs, leurs décisions doivent aussi être prises en temps réel. Pour cela, le processus de collecte des données doit se faire au fur et à mesure que le processus de traitement s'effectue afin de répondre réellement aux besoins des utilisateurs.

2.1.3 Variété : des nouveaux formats de données, plus variés, à intégrer et à analyser

Les Systèmes d'Information classiques exploitent des données structurées d'avance pour stockage et différents traitements. Il s'agit, par exemple, des données de type numérique, caractère, date, image, etc. Mais avec le développement à la fois technologique et des besoins des utilisateurs ; les organisations se sont vu obligées de tenir compte d'autres formats de données dites non-structurées parmi lesquelles nous retrouvons des mails, des commentaires et autres avis issus des réseaux sociaux, des vidéos, etc. La gestion des données massives permet d'intégrer et d'analyser les données à la fois structurées, semi-structurées et non-structurées, collectées depuis des sources diverses.

2.1.4 Véracité : la fiabilité à accorder aux grandes quantités de données

Avec l'augmentation de la quantité, la qualité et la précision se perdent à cause de l'usage des abréviations, des déformations mais aussi des sources peu fiables. Pour y remédier, les données doivent se référer au volume des données existantes qui en principe doivent au préalable être collectées, recoupées, croisées et enrichies pour lever toute incertitude, et ce en respectant le cadre légal afin de créer la confiance et garantir la sécurité et l'intégrité des données.

2.1.5 Valeur : création de la valeur ajoutée pour les entreprises

C'est la dimension la plus importante qui se base sur les quatre premières pour sa concrétisation. Pour que les données puissent avoir de la considération quel que soit leur volume, il est indispensable de les transformer en valeurs exploitables qui vont aider les responsables à

³ Appelé aussi fouille de données, désigne le processus d'analyse de volumes massifs de données sous différents angles afin d'identifier des relations entre les données et de les transformer en informations exploitables.

prendre des bonnes décisions au moment opportun et pourquoi pas anticiper certaines de ces décisions. Ceci revient à dire que les données sans valeur sont inutiles et par conséquent constituent une charge pour l'entreprise. Pour ce faire, l'atteinte des objectifs stratégiques devra être conditionnée par la création de valeur pour tous les consommateurs de services produits par l'entreprise et pour l'entreprise elle-même dans tous les domaines d'activités.

Le respect de ces cinq dimensions par les données repose sur le paradigme MapReduce grâce auquel le Système d'Information est capable de traiter des données structurées et non structurées en provenance de sources diverses (variété) et est capable de traiter un très grand nombre de données en distribuant le traitement sur plusieurs machines (volume, vitesse), lesquels traitements doivent non seulement rendre les données fiables (vérité) mais aussi et surtout donner de la considération (valeur).

2.2 Quelques technologies en vogue pour la manipulation des données massives

Le traitement des données massives (Big Data) s'appuie sur plusieurs technologies. Mais dans cette rédaction, nous allons en présenter deux parmi les plus utilisées, à savoir Apache Hadoop et Apache Spark.

2.2.1 Apache Hadoop

Parti de l'idée originale de Google, Hadoop est une Plateforme open source, développée à l'origine par Yahoo!. Dans ses analyses, Zikopoulos affirme que Hadoop est la Plateforme Big Data par excellence qui se trouve presque systématiquement dans un projet qui vise à traiter de grandes quantités de données.

En effet, de par sa nature, Hadoop s'adresse avant tout à ceux qui ont des volumes de données colossaux d'au moins 10 Téraoctets. En dessous de ce volume, l'utilisation de Hadoop en tant que tel peut apparaître comme contre productive. Pour ce faire, Hadoop contient les éléments ci-après :

1. Hadoop Distributed File System (HDFS)

HDFS est un Système de fichiers distribué, c'est-à-dire que chaque machine composant la grille Hadoop (appelée aussi cluster) va héberger une partie des fichiers. Ce Système de fichiers possède une capacité d'hébergement parfaitement linéaire (chaque octet de disque ajouté à la grille peut participer à l'hébergement des fichiers), une forte tolérance à la panne. Par défaut, chaque fichier est présent en trois exemplaires distribués sur des machines, des racks, voire des data-centers différents. De la même façon, chaque machine apporte ses capacités en entrée-sortie au cluster et la grille est donc capable d'absorber rapidement de grandes quantités d'écriture.

2. Framework Map/Reduce

C'est l'un des composants majeurs de Hadoop, le Framework Map/Reduce. Map/Reduce est un paradigme de programmation permettant de distribuer les algorithmes de traitements au cœur de la donnée. La phase de Map permet de sélectionner et d'organiser les données intéressantes pour le traitement, et la phase de Reduce permet de les agréger, en réalisant par exemple des opérations mathématiques.

Chaque machine du cluster participe aux traitements Map/Reduce, en apportant une fois de plus ses entrées-sorties mais aussi sa CPU (Central Processing Unit). Là encore, la puissance de

calcul disponible est linéaire, et chaque machine ajoutée au cluster ajoutera de la puissance globale.

La distribution des traitements Map/Reduce sur le Système de fichiers lui-même distribué est la grande force de Hadoop. Chaque traitement étant découpé en lot, il suffit d'augmenter la puissance du cluster pour avoir plus de traitements parallèles et donc de réduire les temps de traitement de plusieurs Téraoctets. Dans la mesure du possible, les traitements sont placés là où la donnée se situe, permettant ainsi de profiter au maximum des entrées-sorties d'une machine tout en minimisant l'impact sur le réseau du cluster.

3. Base de données orientée colonnes

Le Framework Map/Reduce fonctionne bien avec les bases de données orientées colonnes. Notons que ce type de base de données a trouvé ses origines chez Google avec le projet BigTable. Le but du projet était de réaliser une base de données accessible par des milliers, voire des millions d'utilisateurs simultanés tout en offrant des possibilités de stockage quasi infinies et des performances adaptées aux exigences du Web moderne.

La solution proposée par Google a été de modifier l'organisation des données sur le disque. Là où les bases de données relationnelles classiques raisonnent en terme de lignes dans une table, BigTable dans sa logique propose de raisonner en terme de colonnes. Ceci permet un modèle de données flexibles (chaque colonne n'est pas forcément présente dans tous les enregistrements d'une table), particulièrement adapté à la modélisation des relations un-à-plusieurs.

4. Machine learning

Le machine learning est une branche de l'intelligence artificielle qui date d'avant le mouvement Big Data, et qui n'y est pas reliée directement. Néanmoins, le machine learning bénéficie d'un fort regain d'intérêt grâce aux nouvelles possibilités offertes par les technologies Big Data.

Le machine learning consiste à construire des Systèmes qui apprennent des données qu'ils reçoivent. Autrement dit, à partir d'un comportement initial, les algorithmes de machine learning sont capables de se corriger et d'évoluer en fonction des faits dont on les alimente.

La construction d'un Système comporte ainsi deux phases. La phase initiale consiste à entraîner la machine, à partir d'un jeu de données connu. Une fois le Système entraîné, il faut le confronter au monde réel et l'évaluer régulièrement afin qu'il se corrige et s'améliore à chaque nouvelle donnée reçue.

5. Emergence du temps réel

Se rapprocher du temps réel est l'un des axes de développement du monde du Big Data. Les Systèmes actuels sont capables de traiter des milliards d'octets de données, mais leur latence est aujourd'hui un véritable enjeu.

Hadoop a posé les standards du monde Big Data, mais son fonctionnement en mode batch l'empêche d'adresser l'ensemble des problématiques de l'industrie du Web. Le temps réel (l'interactivité) est aujourd'hui un facteur différenciant. Chacun des acteurs majeurs du Web l'a bien compris, et tous se concentrent aujourd'hui sur la réduction des temps de traitements.

2.2.2 Apache Spark

Apache Spark est un Framework de traitement Big Data open source construit pour effectuer des analyses sophistiquées et conçu pour la rapidité et la facilité d'utilisation. Celui-ci a originellement été développé par AMPLab, de l'Université UC Berkeley, en 2009 et passé open source sous forme de projet Apache depuis l'année 2010.

Spark présente plusieurs avantages par rapport aux autres technologies Big Data et MapReduce comme Hadoop et Storm. D'abord, Spark propose un Framework complet et unifié pour répondre aux besoins de traitements Big Data pour divers jeux de données, divers par leur nature (texte, graphe, etc.) aussi bien que par le type de source (batch ou flux temps-réel). Ensuite, Spark permet à des applications sur clusters Hadoop d'être exécutées jusqu'à 100 fois plus vite en mémoire, 10 fois plus vite sur disque. Il permet d'écrire rapidement des applications en Java, Scala ou Python et inclut un jeu de plus de 80 opérateurs haut-niveau. De plus, il est possible de l'utiliser de façon interactive pour requêter les données depuis un Shell.

1. Fonctionnalités de Spark

Spark apporte des améliorations à MapReduce grâce à des étapes de shuffle moins coûteuses. Avec le stockage en mémoire et un traitement proche du temps-réel, la performance peut être plusieurs fois plus rapide que d'autres technologies Big Data. Spark supporte également les évaluations paresseuses (lazy evaluation) des requêtes, ce qui aide à l'optimisation des étapes de traitement. Il propose une API de haut-niveau pour une meilleure productivité et un modèle d'architecture cohérent pour les solutions Big Data.

2. Ecosystème de Spark

À côté des API principales de Spark, l'écosystème contient des bibliothèques additionnelles qui permettent de travailler dans le domaine des analyses Big Data et du machine learning. Parmi ces bibliothèques, nous trouvons :

- Spark Streaming : utilisé pour le traitement en temps réel des données en flux. Il s'appuie sur un mode de traitement en « micro batch » et utilise pour les données temps-réel DStream, c'est-à-dire une série de RDD (Resilient Distributed Dataset) ;
- Spark SQL : Permet d'exposer les jeux de données Spark via des API JDBC et d'exécuter des requêtes de type SQL en utilisant les outils BI et de visualisation traditionnels. Spark SQL permet d'extraire, de transformer et de charger des données sous différents formats (JSON, Paquet, base de données) et de les exposer pour des requêtes ad-hoc ;
- Spark MLlib : Est une bibliothèque de machine learning qui contient tous les algorithmes et utilitaires d'apprentissage classiques, comme la classification, la régression, le clustering, le filtrage collaboratif, la réduction de dimensions, en plus des primitives d'optimisation sous-jacentes ;
- Spark GraphX : Est la nouvelle API (en version alpha) pour les traitements de graphes et de parallélisation de graphes. GraphX étend les RDD de Spark en introduisant le Resilient Distributed Dataset Graph, un multi-graphe orienté avec des propriétés attachées aux nœuds et aux arrêtes. De plus, GraphX inclut une collection toujours plus importante d'algorithmes et de builders pour simplifier les tâches d'analyse de graphes.

3. Architecture de Spark

Spark contient trois composants principaux de l'architecture :

- le stockage des données : Spark utilise le Système de fichiers HDFS pour le stockage des données. Il peut fonctionner avec n'importe quelle source de données compatible avec Hadoop, dont HDFS, HBase, Cassandra, MongoDB, etc. ;
- l'API : Permet aux développeurs de créer des applications Spark facilement en utilisant une API standard. L'API existe en Scala, Java et Python. Il existe des liens qui pointent vers les sites présentant les API Spark pour chacun des langages Scala API, Java et Python ;
- le Framework de gestion des ressources : Spark peut être déployé comme un serveur autonome ou sur un autre Framework de traitements distribués comme Mesos ou YARN.

4. Resilient Distributed Datasets (RDD)

Les Resilient Distributed Datasets ou RDD sont un concept au cœur du Framework Spark. Un RDD peut être vu comme une table dans une base de données. Celui-ci peut porter tout type de données et est stocké par Spark sur différentes partitions. Les RDD permettent de réarranger les calculs et d'optimiser le traitement. Ils sont aussi tolérants aux pannes car un RDD sait comment recréer et recalculer son ensemble de données.

En parcourant ces deux technologies Big Data, nous présentons ci-après un tableau comparatif pour en savoir plus.

Tableau 1 : Comparaison des Frameworks Apache Hadoop et Apache Spark

Technologies	Apache Hadoop	Apache Spark
Critères		
Temps de traitement	Vite	100x plus rapide en mémoire et 10x plus rapide sur disque que Hadoop
Mode de traitement	Par lot	En temps réel et par lot
Stockage des données	Sur disque	En mémoire et sur disque
Langage de développement	Java	Scala
Langage supporté	Java	Scala, Java, Python, Clojure, R
Système d'exploitation supporté	Linux	Linux, Windows, Mac OS

De ce tableau précédent, il ressort clairement qu'Apache Spark a été développé non pas pour remplacer Apache Hadoop, mais plutôt pour le compléter afin de gagner davantage du temps dans le traitement des données ; ce qui constitue un élément de taille sinon primordial, étant donné le volume de données qui doivent être manipulées par les moteurs de traitement Big Data.

2.2 Population et échantillon

La Haute Ecole de Commerce de Kinshasa est une institution publique de formation régie par l'ordonnance n° 71-075 du 06 août 1971 et située sur l'avenue de libération (ex. 24 novembre) dans la commune de la Gombe. Elle constitue notre champ d'investigation, en raison non seulement du nombre d'étudiants qu'elle comporte (29.971 pour l'année académique 2023-2024), mais aussi pour la pluralité de spécialités organisées⁴.

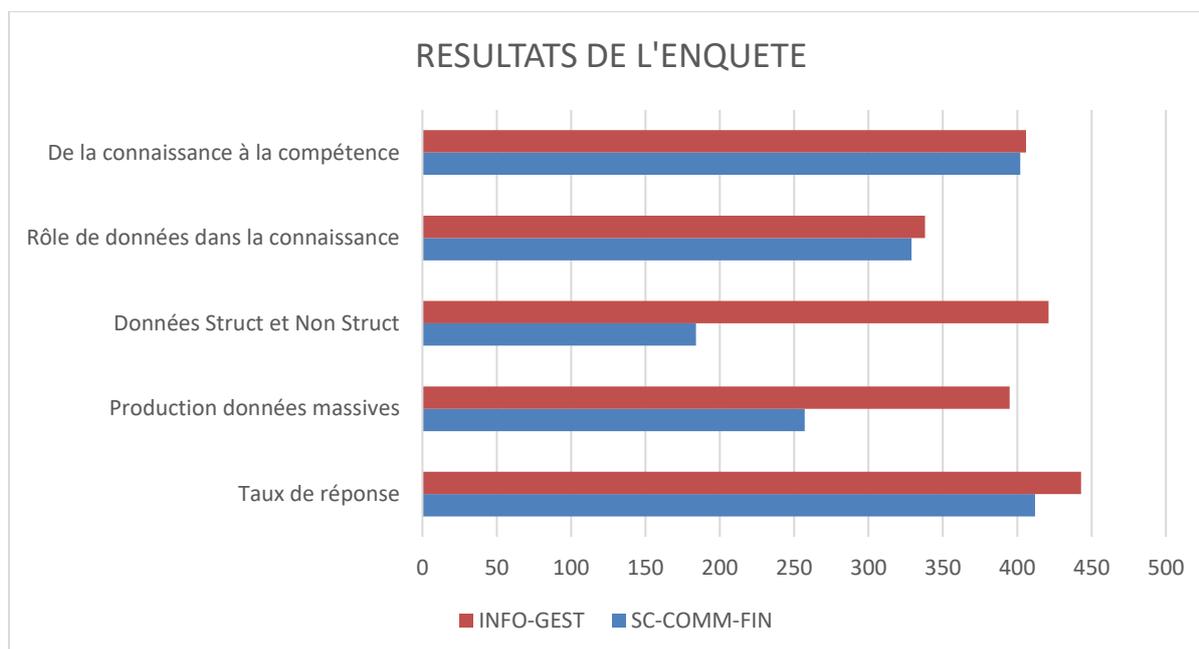
⁴ <https://heckin.ac.cd>

La HEC est un établissement pilote de tous les ISC⁵ de la République Démocratique du Congo, et de ce fait, constitue le miroir de tous ces établissements. Ce qui revient à noter que l'échantillon choisi sera représentatif de tous les étudiants disséminés à travers le pays dans le domaine de gestion.

Nous avons choisi les deux grandes sections en terme d'effectifs à savoir la Section Sciences Commerciales et Financières et la Section Informatique de Gestion, et avons opté pour un échantillon de 1000 étudiants, en raison de 500 étudiants par section et indépendamment de promotions.

2.3 Résultats

Graphe 1 : Résultats de l'enquête



2.4 Discussion

Les données relatives à l'enquête effectuée au sein de la Haute Ecole de Commerce de Kinshasa dans les deux principales sections en terme d'effectifs renseignent, en ce qui concerne le taux de réponse au questionnaire administré, que la majorité absolue des enquêtés a répondu. Ceci est dû au fait, non seulement de l'environnement immédiat dans lequel nous évoluons, mais aussi parce que le questionnaire était rendu disponible, en plus de la version papier, en version mobile, étant donné le taux d'utilisation des appareils mobiles de nos jours dans la société. D'où les taux de réponse respectivement de 82,4 % et 88,6 dans la Section Sciences Commerciales et Financières et la Section Informatique de Gestion. Ce qui nous encourage à analyser les résultats issus de cette enquête.

A la question de savoir si les enquêtés ont connaissance du fait qu'ils produisent chacun des données massives, la tendance s'est beaucoup plus penchée du côté des étudiants informaticiens avec 89,1% contre 62,3% pour leurs collègues financiers. A notre avis, nous pensons que cet écart ne devrait pas être important parce que tous les enquêtés, quel que soit leur domaine d'apprentissage, ont accès aux ordinateurs et possèdent tous des appareils mobiles qui leur

⁵ Institut Supérieur de Commerce de Kinshasa, actuelle Haute Ecole de Commerce de Kinshasa

permettent, grâce aux réseaux sociaux, à la fois de produire et de consommer des données en grande quantité. Il suffit d'accéder à leurs abonnements en forfaits Internet pour se rendre compte de cette réalité. Qu'à cela ne tienne, nous avons, grâce au questionnaire leur administré, expliqué le terme « données massives » ou « big data », pour que chacun puisse avoir une idée sur la réponse à donner. D'autre part, nous comprenons aussi cet écart dû au fait que les étudiants informaticiens ont reçu abondamment d'explications dans le cadre de leur formation. Toutefois, nous pensons que cet écart ne sera plus consistant à l'issue de cette étude.

Quant à la différence à ressortir entre les données structurées et celles non structurées, l'écart entre les deux catégories, 95% pour les informaticiens contre 44,7% pour les financiers, est justifié du fait que les premiers, non seulement ont des cours traitant des bases de données relationnelles⁶, mais aussi développent des applications en recourant à la fois aux bases de données relationnelles et aux bases de données NoSQL⁷, sans oublier des bases de données Web et mobiles. Nous insistons sur le fait que, quel que soit le domaine de prédilection pour chacun des enquêtés, la connaissance et le recours au domaine informatique est plus qu'une nécessité de nos jours, pour quiconque voulant se maintenir dans son domaine tout en faisant face à la concurrence. Donc le domaine ne doit pas être présenté comme une excuse pour justifier ce faible taux, car l'Internet et toutes ses applications servent aussi à former et à se faire former dans tous les domaines de la vie humaine.

Il est important de savoir que les données sont considérées comme la matière première sur laquelle tout système d'information doit se pencher pour traitement afin de donner des livrables appelés informations. Et les décideurs doivent se pencher sur ces dernières pour avoir de la connaissance dans leurs domaines d'activités afin de prendre les différentes décisions avec compétence. A la question de savoir le rôle que jouent les données dans le processus de connaissance, les deux catégories ont présenté presque le même taux, 76,3% pour les informaticiens et 79,9% pour les financiers. Nous interprétons ce score par le fait que les deux catégories recourent aux données dans le cadre de leurs activités, à savoir les données comptables et les données informatiques. A ce stade il faut retenir qu'en amont, toute donnée qu'elle soit manuelle ou numérique doit contenir de la valeur ajoutée pour qu'en aval l'information produite soit de qualité, c'est-à-dire celle qui doit permettre au décideur d'être compétent dans la prise de décisions.

Concernant la dernière question relative à la compétence qui doit être conséquente au degré de connaissance ; une fois de plus, les deux catégories ont présenté presque le même taux (91,6% pour les informaticiens et 97,6% pour les financiers), pour la simple raison que cette question fait suite à la précédente. Comme c'est un processus qui doit partir des données de valeur et que ces dernières sont exploitées dans les deux catégories à une différence près, nous ne pouvons que comprendre ce résultat. Toutefois, nous insistons sur le fait que les données numériques permettent d'avoir accès à une compétence exigée en cette période de véritables mutations, où les entreprises et les utilisateurs évoluent dans un monde marqué par la concurrence sur tous les plans. Douç l'importance de se maintenir dans le présent et avoir de vision pour l'avenir en toute compétence.

⁶ qui permettent la création des données structurées.

⁷ Not Only SQL : permettant la création des bases de données structurées et non structurées, afin de répondre au principe de Big data.

3. Conclusion

De ce qui précède, nous retenons que les technologies comme les besoins des entreprises et des utilisateurs ont pris une autre dimension, celle de produire et de consommer de quantités énormes de données à travers des dispositifs tels que les ordinateurs simples, le réseau Internet ou encore les appareils mobiles.

Il a été question à travers cet article d'expliquer les notions que tout le monde exploite, mais sans en connaître le bienfondé, à savoir les données massives ou big data. Ces dernières ont poussé les uns et les autres à devenir budgétivores en matière de production et de consommation de données. Nous avons insisté et démontré que les données, quelle que soit leur quantité, doivent contenir de la valeur ajoutée pour que, une fois traitées, puissent produire des informations de qualité, afin que ceux qui doivent prendre des décisions puissent être compétents, car possédant de la connaissance avérée dans leurs domaines respectifs.

L'enquête effectuée a démontré qu'il est possible d'atteindre ses objectifs en toute compétence dans son domaine, si et seulement si les données utilisées en amont en dépit de leurs natures et quantités, possèdent de la valeur ajoutée.

Références Bibliographiques

- [1] Monino. J.L. (2021, feb.). « Big Data, Open Data et Valorisation des données. Réseau de Recherche sur l'innovation », Document de Travail, n° 48, 2017, pp. 13-14. [En ligne]. Disponible : <http://rrifr.univ-littoral.fr/wp-content/uploads/2015/10/doc48.pdf>
- [2] Mukenge Mbumba J. (2023). Quel Avenir Pour Les Systèmes D'Information De Gestion Traditionnels ?, Editions Universitaires Européennes, OmniScriptum.
- [3] Vaughan J. (2021, aug.). « Spark et Hadoop : compagnons ou concurrents ? », in *Le Mag IT.* [En Ligne]. Disponible : <http://www.lemagit.fr/conseil/Spark-et-Hadoop-compagnons-ou-concurrents>
- [4] Yann G. (2013). *Le management visuel de l'information à l'heure du Big Data*, Infodesign, Paris : Ellipses.