Revue maghrébine de documentation et d'information, n°30, 2021

L'archivage à l'ère des Big Data : quels nouveaux paradigmes pour l'Archive et quels rôles pour l'archiviste ?

Archiving in the era of Big Data: What new paradigms for archives and what is the roles of the archivist ?

الأرشيف في عهد البيانات الضخمة :
أي نموذج جديد للأرشيف وأي دور للأرشيفي ؟

Reine Marie Ndéla Marone

EBAD - Université Cheikh Anta Diop de Dakar, SENEGAL

Laboratoire de recherche en science de l'information et de la communication (LARSIC),

Enseignant-chercheur en Informatique documentaire à l'EBAD et membre du LARSIC, je m'intéresse à l'informatique et à ses applications notamment dans le champ des sciences de l'information et la documentation.

reinemarie.marone@ucad.edu.sn

Moustapha Mbengue

EBAD - Université Cheikh Anta Diop de Dakar, SENEGAL

Laboratoire de recherche en science de l'information et de la communication (LARSIC),

Enseignant-chercheur en SIC, directeur de l'EBAD, membre fondateur du LARSIC. S'intéresse à la culture numérique, aux systèmes d'information Web à l'informatique documentaire et au comportement informationnel.

moustapha.mbengue@ucad.edu.sn

Résumé : Cet article tente de voir dans quelle mesure les technologies peuvent aider à archiver et à exploiter le flux grossissant de données numériques. L'étude s'inscrit dans le cadre théorique de l'archivage des Big data et elle utilise la méthodologie de la recherche et de l'analyse documentaire. Les résultats de l'étude montrent que l'archiviste peut jouer un rôle très important dans la gestion des Big data, qu'il peut veiller à ce que les données soient placées dans leur contexte afin de conserver leur sens métier et aider ainsi à résoudre le problème du vrac informationnel. L'archiviste peut également permettre d'intégrer à un socle technique, des méthodes d'archivage qui constitueront alors une couche fonctionnelle permettant une meilleure gestion du cycle de vie des données dans le temps.

Mots-clés : Big Data, Archives, Apache Hadoop, Apache Spark, GED.

Abstract : This article attempts to view how technologies can help to archive and to exploit the growing flux of digital data. The study is part of the theoretical context of big data archiving. It uses the methodology of research and literature review. The results of the study show that archivist can play a vital role in the management of big data. An archivist can indeed ensure that data are placed in their context in order to maintain their meaning and thus helps to solve the problem of informational bulk. The archivist can also make it possible to integrate archiving methods into a technical base, which will then constitute a functional layer allowing better management of the data life cycle over time.

Keywords : Big Data, Archives, Apache Hadoop, Apache Spark, EDM.

مستخلص : يحاول هذا المقال معرفة الى أي مدى تستطيع التكنولوجيا أن تساعد على أرشفة واستغلال هذا المدّ المتزايد من البيانات الرقمية. تندرج الدراسة ضمن الإطار النظري لأرشيف البيانات الضخمة، وتعتمد على منهجية البحث والتحليل الوثائقي. أظهرت النتائج أن الأرشيفي يستطيع القيام بدور هام في إدارة البيانات الضخمة وبوسعه أن يضعها في سياقها حتى تحافظ على معناها. كما أن له من المهارات ما يجعله قادرا على حل مشكلة تشتت المعلومات، وإدراج طرق الأرشفة في الجهاز التكنولوجي لأجل أن تكون وظيفية ومساعدة على الإدارة الجيدة للدورة العمرية للبيانات.

الكلمات المفتاحية : البيانات الضخمة، الأرشيف، الإدارة الإلكترونية للوثائق.

Introduction

Les Big data ou mégadonnées ou encore données massives désignent des données numériques qui peuvent être de nature variées (textes, photos, sons, vidéos, graphiques, chiffres, liens hypertextes, clics, pages vues…) et qui sont produites en masse par l'activité quotidienne des entreprises privées ou publiques, ou des organismes scientifiques. Ces données sont issues entre autres de l'open data (libéralisation des données publiques), des objets connectés, des réseaux sociaux, des données de la recherche… Le Big data a impacté plusieurs domaines y compris celui des sciences de l'information et de la communication, l'archivistique en particulier. En effet les systèmes traditionnels de gestion documentaire utilisés encore dans le monde des archivistes ne sont pas forcément adaptés à la gestion des Big data. Le traitement de ces derniers nécessite des outils et technologies spécifiques, pour pouvoir en tirer profit. C'est pour cela qu'au cours des dernières années, les fournisseurs de technologies ont introduit de nouvelles technologies telles que celles d'analyse et de traitement des Big data qui permettent aux organisations d'exploiter la valeur de volumes massifs de données.

Selon Bastin et Tubaro (2018) « Le Big data fait référence à des données massives variées, qui ne sont pas appréhendables par l'intuition et les capacités humaines d'analyse, mais également celles des systèmes informatiques classiques, tant leur volume est important ». L'apparition du vocable « Big data » remonte à la fin des années 90 lorsque des chercheurs ont remarqué que : les capacités de stockage et de calcul qui augmentent sans cesse (loi Moore) ont entraîné un volume de données générées qui dépasse les facultés de représentations humaines (Delort, 2015). Il faut également noter le fait que l'essor d'internet a eu un effet multiplicateur sur la génération de données à l'échelle mondiale.

Selon Espinasse et Bellot (2017) « Les données qui constituent le Big data sont de nature et de provenance très diverses : elles peuvent être non structurées ou semi structurées et générées par des sources multiples ». Ces données peuvent s'avérer utiles pour de nombreux acteurs de la société comme les entreprises et les gouvernements. Pour traiter ces données, des technologies avancées adaptées sont utilisées, qui permettent par exemple d'élaborer des diagnostics, de prendre des décisions en conséquence et d'établir des plans d'action. Parmi ces technologies on peut citer celles de l'intelligence artificielle et du machine learning qui peuvent être utilisées à plusieurs fins, par exemple elles peuvent servir à explorer les données afin de détecter des tendances et des régularités dans le comportement des consommateurs en ligne, ou

encore du trafic routier (Bastin et Tubaro, 2018 ; Seguy, 2019 ; Charlin, 2017) … Ces technologies sont utilisées par des entreprises de marketing pour analyser les habitudes de consommation de leurs clients et établir des publicités ciblées (personnalisées). C'est le cas la plate-forme vidéo Netflix qui analyse les données de visionnage pour évaluer les préférences de ses abonnés et déterminer ainsi des tendances. Grâce à ces informations, Netflix peut recommander le bon contenu à l'utilisateur, mais aussi produire le film qui trouvera facilement son public (Seguy, 2019). La Blockchain fait également partie de ces technologies qui peuvent être utilisées pour faire des traitements dans les mégadonnées. Elle permet par exemple de gérer la traçabilité des documents, les horodater et conserver l'ensemble des accès qui ont été effectués sur ces derniers (Mougayar, 2016).

Le Big data a ainsi touché tous les secteurs, notamment l'archivistique et la bibliothéconomie qui sont aujourd'hui confrontées aux problèmes de l'organisation et de la mise en valeur de l'information face à la massification des données (Laurent et al., 2015 ; Vieira, 2018 ; Ben Amor et Mkadmi, 2018). L'explosion des volumes de documents, mais aussi des données et métadonnées n'est plus gérable par les systèmes de gestion documentaire classiques tels que la GED (Gestion Électronique des Documents), les SAE (Systèmes d'Archivage électronique), nativement orientés vers les bases de données relationnelles, utilisées encore dans ces secteurs (Vieira, 2018 ; Phillipon et Besson, 2020 ; Goupil, 2017). Les systèmes de gestion documentaire traditionnelle sont de ce fait confrontés à de nouveaux challenges auxquels le professionnel de l'information documentaire doit répondre afin de pouvoir tirer de la valeur (ou de l'information) des grandes masses de données archivées (Vieira, 2018 ; Goupil, 2017).

Il convient dès lors de se demander si les technologies du Big data peuvent aider les systèmes de gestion documentaire traditionnels à répondre aux défis de l'archivage des grandes masses de données ? Aussi, quelles nouvelles missions l'archivage des Big data donne-t-il à l'archiviste ? Autant de questions que soulève la problématique de l'archivage des mégadonnées.

Nous formulons l'hypothèse qu'un changement du socle technologique traditionnel fourni par la plupart des systèmes d'archivage traditionnel s'impose pour faire face aux données massives et que les nouvelles technologies du Big data peuvent aider à relever le défi de l'archivage des mégadonnées et permettre aux organisations d'exploiter leurs données documentaires de manière plus efficace et à moindre coût.

De même les Big data et les technologies permettant de les traiter valorisent l'archiviste en lui permettant de structurer au mieux les données afin de permettre l'analyse, les calculs sur les données et non plus seulement pour trouver les documents unitaires, mais également de veiller à ce qu'elles soient placées dans leur contexte. L'archiviste peut également permettre d'intégrer à un socle technique, des méthodes d'archivage qui constitueront alors une couche fonctionnelle permettant une meilleure gestion du cycle de vie des données dans le temps.

La première partie de notre étude est une introduction qui présente le contexte de l'archivage dans les Big data, la deuxième partie est consacrée au cadre théorique et conceptuel et la troisième partie présente l'étude des outils et méthodes pour l'archivage des mégadonnées, elle est suivie d'une discussion sur les limites possibles de ces technologies et du rôle que l'archiviste peut jouer pour aider à combler ces limites.

1. Cadre conceptuel et théorique

Notre étude s'inscrit dans le cadre théorique de l'archivage des Big data, mais il nous paraît tout d'abord nécessaire de faire la différence entre l'archivage et le stockage de données. En effet le débat sur le stockage et l'archivage date de plusieurs années et il est courant de voir une confusion entre l'archivage et le stockage de données (Ben Amor et Mkadmi, 2018). Dans leur article « Et demain ? Archivage et Big data » Marie-Odile Charaudeau et al. (2015) tentent de donner une explication pour cerner la différence entre le « big archivage » et le « big stockage », ils soutiennent que « *l'archivage et le stockage sont différents et précisent que le choix d'un système de stockage ou d'archivage dépend des objectifs que l'on recherche* ». Paul Haquette (2019) dans son article « *Le big data libère le potentiel des archives d'entreprises* » va encore plus loin en affirmant que « *l'archivage de données permet de conserver pour la conformité réglementaire et indexer pour faciliter leur localisation et leur récupération, les anciennes informations qui ne sont pas nécessaires pour les opérations quotidiennes, mais qui sont encore nécessaires et importantes pour références dans le futur. Alors que le stockage de données quant à lui se résume à copier les fichiers en vue de les restaurer si le besoin s'en fait sentir, par exemple après un sinistre ou pour restaurer les fichiers perdus ou corrompus* ». L'archivage se distingue aussi du stockage par la notion de sécurité qui les différencie. Les données archivées contrairement à celles stockées ne peuvent pas être modifiées, il n'est pas possible de les détruire à moins d'effectuer un contrôle minutieux, et toute opération effectuée doit être répertoriée en vue d'être retracée (Haquette, 2019). Les experts en stockage affirment

constamment que : « les sauvegardes ne sont pas des archives » (Ben Amor et Mkadmi, 2018). D'ailleurs, les logiciels de stockage traditionnels n'aideront pas à archiver, les fournisseurs de logiciels de stockage et de récupération de données ont commencé à intégrer différentes fonctionnalités dans leurs logiciels comme la déduplication des données et la gestion du cycle de vie des données avec la hiérarchisation du stockage (Ben Amor et Mkadmi, 2018).

Cependant même si l'archivage consiste à conserver les données pour référence dans le futur. Comme le précise Michèle Battisti (2015), il ne s'agit pas uniquement de conserver les données et les mettre à part, mais au contraire, de les laisser circuler : « *conserver des données numériques, c'est les laisser circuler et les faire migrer continuellement d'un serveur à l'autre par leur découpage en morceaux* ». L'archivage ne consiste donc pas à fixer les données sur un support, mais plutôt à les préserver pour leur valeur historique, probante ou tout simplement leur degré de criticité. L'archivage est également un processus dynamique qui précède les connaissances : l'archive « *ne suit plus la connaissance, mais la précède* » (Battisti, 2015). On pense et réfléchit alors au processus d'archivage avant même que les données ne soient produites et interprétées ultérieurement.

La circulation des données n'est cependant pas sans conséquence sur leur mode d'archivage, car, les données deviennent de plus en plus volumineuses et se déplacent trop rapidement, mais également ne correspondent pas aux structures des architectures de base de données relationnelles actuelles (Vieira, 2018). À cela s'ajoutent les besoins de stockage, de reconnaissance, de contrôle, d'archivage légal à valeur probante, de recherche ainsi que le besoin d'accès direct par les utilisateurs concernés qui ne font que croître. Tous ces faits challengent les solutions historiques d'archivage et de plus en plus d'entreprises s'interrogent sur le devenir de leur existant souvent « progicialisé » ou externalisé (GED-as-A-service), dont les coûts sont historiquement jugés hors de proportions par rapport aux attentes opérationnelles ou pires, ne délivrent pas les fonctions attendues (Phillipon et Besson, 2020). Les systèmes actuels de type EDM (Electronic Document Management) / DAM (Digital Asset Management) / ECM (Enterprise Content Management) ne sont en effet pas forcément calibrés pour gérer un tel volume et une telle variété de données. Il est nécessaire et fondamental selon Phillipon et Besson (2020) « que ces systèmes gagnent en flexibilité, modularité, performance et fonctionnalités pour répondre à ces nouvelles exigences opérationnelles et réglementaires ». C'est donc l'occasion de faire un point sur les stratégies d'archivage possibles pour

trouver des solutions à ces problématiques en tirant parti des nouvelles technologies et approches (Goupil, 2017), sachant par ailleurs que la nature des documents à gérer par l'entreprise risque de profondément évoluer encore dans les années à venir, notamment en termes de preuve de véracité (identité, situation patrimoniale, adresse…) ainsi que leur variété (conversation robotisée, empreinte de reconnaissance faciale…) (Phillipon et Besson, 2020). Cette variété et évolution des données d'archivage est d'ailleurs évoquée sous le nom de « big archive » par des auteurs comme Mariam Esteva et al. (2013) dans un article intitulé « Data Mining for Big Archives Analysis : a Case Study » ou Martyn Joll (2009) dans son article « Big Archives and Small Collections ».Ces auteurs soulignent la variété des données en termes de contenu, de type et de modèle d'organisation, mais également leur volume qui devient de plus en plus important. Qu'ils soient nativement numériques ou numérisés, les archives sont de plus en plus importantes en termes de quantité et deviennent de plus en plus complexes à gérer, nécessitant ainsi de nouvelles méthodes de calcul pour les traiter d'une manière efficace.

Les technologies Big data ouvrent donc de nouvelles possibilités pour l'archivage des données (Ben Amor et Mkadmi, 2018), « en tirant parti des normes ouvertes, des plates-formes d'intégration, de consolidation et de montée en puissance » (Goupil, 2017).

Les Big data représentent une révolution dans le domaine des SIC et elles ont envahi plusieurs secteurs d'activités parmi lesquels on peut compter l'archivistique. On peut ainsi observer une préoccupation croissante dans la littérature archivistique caractérisant les archives numériques comme des mégadonnées (Mordell, 2019). Ainsi le couplage entre archives et mégadonnées a fait l'objet de plusieurs publications dont on peut tirer la conclusion suivante : les technologies Big data ouvrent de nouvelles possibilités pour l'archivage des données (Ben Amor et Mkadmi, 2018 ; Goupil, 2017). En effet les archives sont de plus en plus importantes en termes de quantité et deviennent de plus en plus complexes nécessitant alors de nouvelles méthodes et techniques de calcul pour les traiter d'une manière efficace (Ben Amor et Mkadmi, 2018 ; Esteva et al., 2013). Ce qui invite donc à revoir les stratégies d'archivage en tirant parti des nouvelles approches et technologies existantes.

2. Présentation des outils

Comme souligné par Forrester (2015), les technologies Big data ouvrent de nouvelles possibilités à l'archivage des données, « *en tirant*

parti des normes ouvertes, de l'intégration de plates-formes évolutives » (Yuhanna et al., 2015). Apache Hadoop/Spark et les bases de données NoSQL sont en ce sens, des plateformes évolutives pouvant stocker et traiter de très gros volumes de données structurées, non structurées et semi structurées. Au-delà du stockage d'autres fonctionnalités telles que la recherche, la découverte et l'exploration, pour la conformité et l'analyse nous semblent pouvoir aussi aider la GED et les SAE traditionnels à répondre aux nouveaux enjeux de gestion des mégadonnées, auxquels ils sont confrontés.

Eu égard à leur taille et à leur vitesse de croissance, le traitement des Big data nécessite des outils spécifiques pour pouvoir en tirer un intérêt économique, historique et social. Leur traitement et leur conservation nécessitent des processeurs rapides, de grandes capacités de mémoire vive, de nouvelles méthodes et techniques de traitement, d'analyse et de gestion de flux particuliers.

Il convient dès lors, d'apporter des éléments de réponses à la question de l'archivage des Big Data notamment en identifiant des outils, des moyens et des approches pouvant permettre de traiter et d'archiver à l'ère du Big data, des données dont le flux ne cesse d'augmenter.

2.1. Intérêt des bases de données NoSQL pour l'archivage des Big data

Les bases de données relationnelles et SGBDR (Système de Gestion de Bases de données relationnelles) constituent le modèle d'architecture le plus utilisé encore dans l'architecture des systèmes de gestion documentaire traditionnels (Vieira, 2018). Les SGBDR les plus connus sur le marché sont MySQL, Oracle, SQL SERVER et PostgreSQL. Les bases de données relationnelles ne sont toutefois pas adaptées aux besoins du Big data (Vieira, 2018).

2.1.1. Bases de données Relationnelles

Une base de données relationnelle (BDR) est constituée de plusieurs tables ou relations manipulées grâce à des opérations tels que l'intersection, l'indexation, la jointure, le produit cartésien, etc. Elle stocke les données actives, interrogées souvent et à une grande fréquence et la plupart du temps par un grand nombre d'utilisateurs.

Les systèmes documentaires traditionnels tels que les solutions de GED et SAE classiques utilisent des bases de données de type relationnelles.

2.1.2. Datawarehouse

Il existe également une autre catégorie de bases de données dédiées souvent installées, appelées datawarehouses, qui synthétisent les informations et permettent de répondre à des besoins précis. Un datawarehouse regroupe des données hétérogènes comme des fichiers Excel, texte, XML…, issues de multiples applications du SGBDR. Le datawarehouse permet l'historisation de la donnée en conservant pour chaque donnée : une **date** ou un **numéro de version**, afin d'éviter les doublons par rapport à la base de données, et ainsi obtenir une certaine traçabilité.

Toutefois, les datawarehouses peuvent être coûteux pour le Système d'information, car les bases de données stockent difficilement les très grands volumes de données, et le temps de traitement des données entrantes peut être très long (à cause de l'indexation). De plus, ils ne prennent pas en compte les données non structurées : journaux web non structurés ou graphiques point à point…

Ainsi les systèmes documentaires tels que les solutions de GED et SAE qui utilisent traditionnellement des bases de données de type relationnelles sont aujourd'hui en train de migrer petit à petit vers celles NoSQL (non relationnelles) afin de faciliter le stockage en grande quantité et surtout la rapidité de traitement des mégadonnées (Vieira, 2018).

2.1.3. Bases de données NoSQL

Les bases données NoSQL permettent le stockage et le traitement de données structurées et non structurées, de façon distribuée dans des clusters de serveur à bas coût, liés entre eux par plusieurs milliers de nœuds. Les bases de données NoSQL rompent avec les principes des bases relationnelles, car une grande majorité d'entre elles sont dépourvues de tout schéma ou modèle de base de données construit en préalable (il n'y a pas de jointure entre les tables par exemple) ce qui les rend hautement dynamiques et leur permettent de stocker et de traiter des données peu ou non structurées. Un autre atout des bases NoSQL est qu'elles sont open source ce qui leur permet d'obtenir une forte interopérabilité au sein des SI (Système d'Information).

Plusieurs solutions de type NoSQL sont proposées sur le marché actuel, chacun d'eux ayant sa propre orientation et une structure différente. Il y a par exemple les bases de données orientées agrégats (BDOA) qui remplacent par le concept d'agrégat, la notion de table (propre aux SGBD classiques) et qui permettent d'y stocker les données

les plus sollicitées (Espinasse et Bellot, 2017).Ces types de bases de données peuvent être scindées en 3 catégories qui sont : les entrepôts clé – valeur, les bases de données orientées documents (BDOD), les bases de données orientées colonnes. Dans le milieu des spécialistes de l'information, les bases de données NoSQL orientées documents (BDOD) sont d'ailleurs bien connues, car elles peuvent se connecter à des solutions ECM (Enterprise Content Management) telles que Sharepoint ou Alfresco, c'est-à-dire des applications qui manipulent naturellement des documents (Lemberger et al., 2016).

Le fonctionnement des bases de données NoSQL orientées document est basé sur le couple clé — valeur. Les valeurs peuvent être des objets semi-structurés ou non structurés, par exemple des documents en général au format XML ou JSON, ou même PDF ou Microsoft Office (MS Word, Excel, etc. …). Dans la BDOD, les documents peuvent être de format différent tout en étant situés sur le même emplacement, c'est-à-dire plusieurs documents qui possèdent chacun une structure différente, peuvent être stockés dans une même table. De plus, il n'y a pas de modèle prédéfini, ce qui veut dire qu'il est possible de modifier la structure d'un document à la volée. Avec un SGBDR tout cela aurait été beaucoup plus complexe à gérer, ce qui signifie donc que les bases de données non relationnelles sont plus malléables et dynamiques, car leurs champs s'adaptent à différentes structures de documents. Ainsi, les professionnels de l'information documentaire peuvent profiter de la souplesse des bases de données non relationnelles face à des situations de « Big data », notamment pour des fonctionnalités métiers destinées à la gestion électronique des documents.

La figure suivante (figure 1) représente une base de données orientée document (BDOD) pour une bibliothèque.

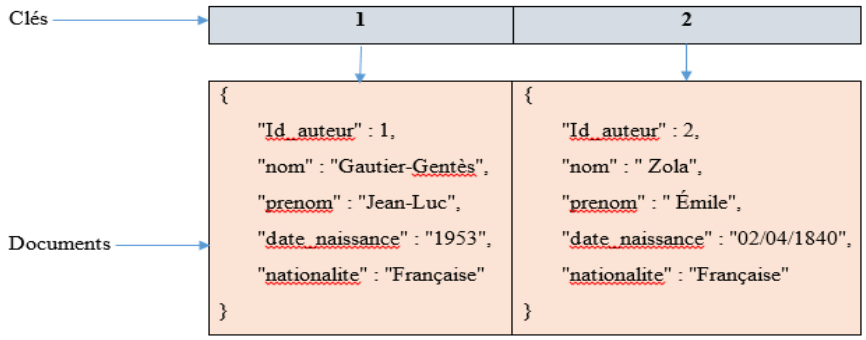

Figure 1 : Base de données orientée document (BDOD)

Dans cet exemple, chaque auteur est représenté par une clé et un document contenant les informations le concernant.

Toutefois pour soutenir cette gestion des grandes masses de données, il est important de déployer une infrastructure de stockage approprié tel que le framework Apache Hadoop.

2.2. Framework Apache Hadoop

Hadoop est un framework Java libre qui permet la mémorisation de grandes quantités de données (« Big data »), la répartition et le traitement parallèle de ces données dans des environnements distribués horizontalement (Espinasse et Bellot, 2017). Le déploiement horizontal (« Scale out ») représente l'un des piliers de Hadoop, car il permet de répartir de grandes quantités de données sur plusieurs milliers de serveurs qui se trouvent dans un cluster d'ordinateurs, de façon fiable et en bénéficiant d'une tolérance de panne. Il ne s'agit donc plus de mettre en œuvre des serveurs spéciaux et coûteux (« Scale up »), mais d'investir sur du matériel courant, d'un coût bien plus abordable. Dans les environnements Big data, Hadoop est considéré comme un standard pour les applications Big data (Espinasse et Bellot, 2017).

Dans Hadoop les données sont stockées sous forme d'une multitude de fichiers distribués qui sont regroupés et structurés lors de la phase d'analyse. Cette méthode de conservation est recommandée pour de gros volumes de données qui ne nécessitent pas une structure analytique a priori, c'est le rôle du « datawarehouse » qui reste la structure la mieux adaptée à l'analyse répétitive et comparative des données structurées de l'entreprise (Nieuwbourg, 2017).

Hadoop est l'implémentation la plus populaire et la plus mature sur le marché, du modèle MapReduce (Espinasse et Bellot, 2017). MapReduce est un modèle de programmation créé par Google pour le traitement et la génération de grands ensembles de données sur des clusters d'ordinateurs. Le framework MapReduce propose deux étapes : Map et Reduce pour effectuer un travail. L'étape Map consiste à répartir le travail sur les différents nœuds du cluster d'ordinateurs. Par la suite, l'un des nœuds appelé le **nœud master** ou **name node**, combine les résultats fournis par chaque nœud, en une seule réponse cohérente à une requête : il s'agit de l'étape Reduce. Cela est rendu possible grâce à son système de fichiers distribués : HDFS. Étant donné que dans l'environnement des Big data, une seule machine ne peut pas supporter la quantité de données présente, le système de fichiers distribués HDFS est mis en place sur tous les nœuds du réseau. Le framework fractionne les

données en gros blocs et les distribue à travers les nœuds d'un cluster d'ordinateurs. Ensuite pour traiter les données, il distribue le travail entre les différents nœuds appelés **nœuds esclaves** ou **data nodes** et chaque nœud traite les données qui lui sont attribuées. De cette manière les données sont traitées plus rapidement et plus efficacement que dans une architecture plus classique.

Une autre machine, appelée **secondary name node** sert de name node de secours en cas de défaillance du nœud maître (nœud master) et il a donc pour rôle de faire des sauvegardes régulières des données.

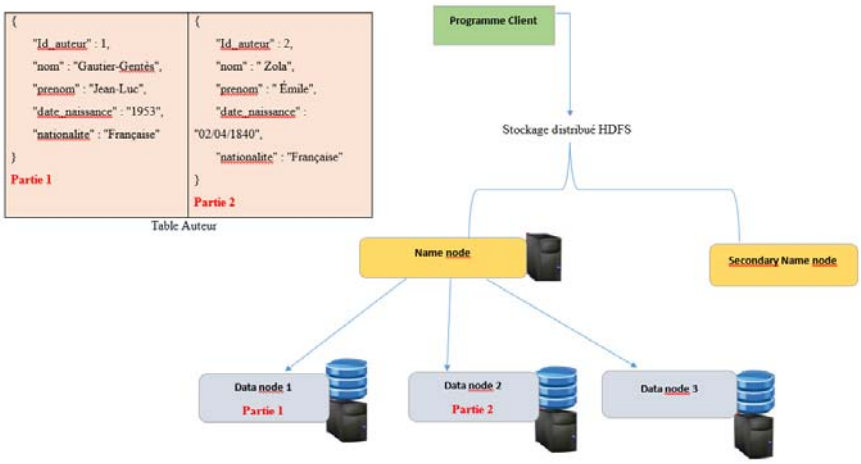

Figure 2 : Fonctionnement d'un programme utilisant Hadoop

La figure 2 présente le fonctionnement d'un programme qui utilise Hadoop en combinaison avec une base de données NoSQL (base de données d'une bibliothèque que nous avons vu à la section 3.1.3). Ce programme peut être un système documentaire tel qu'une GED ou même un SAE.

Dans cette figure, la table contenant l'ensemble des auteurs est subdivisée en plusieurs fichiers (Partie 1 et Partie 2 dans cet exemple). Et chaque fichier est transféré et stocké dans un nœud esclave (data node). De même les tâches sont distribuées entre ces nœuds esclaves : les data node (Data node 1, Data node 2…). Par la suite, le nœud master ou name node, combine les résultats fournis par chaque nœud esclave, en une seule réponse cohérente à une requête.

Ainsi lorsqu'un utilisateur fait une requête telle que la demande d'une page ou tout autre type de traitement au sein du système

documentaire, il utilise un programme client installé sur son poste de travail, qui peut un être son navigateur. Le traitement à faire est divisé alors en plusieurs tâches distribuées au sein des ordinateurs sur le cluster (**data node)**. Chaque data node effectue les traitements qui lui sont alloués, sur les fichiers mis à sa disposition. Il transmet ensuite ses résultats au **name node** qui les combine et envoi le résultat final au programme client. Ce programme client les affiche alors à l'utilisateur.

Hadoop dispose également d'outils qui permettent de le connecter avec d'autres applications. Sqoop est une technologie clé qui fait partie de cette gamme d'outils et il est utilisé pour faire l'archivage. Il permet de déplacer les données à archiver d'un SGBD relationnelle vers Hadoop. Ainsi on peut transférer des données situées dans des SGBD classiques vers une application d'archivage qui utilise Hadoop. Ce transfert permet de rendre la donnée archivée rapidement disponible, car les données ne sont plus stockées sur un support annexe (cassettes, espace de stockage indépendant), mais c'est le HDFS d'Hadoop qui va se charger d'assurer leur stockage directement au même niveau que les données les plus actives et ce de façon efficace et sécurisée.

Un autre avantage d'Hadoop est qu'on peut étendre un cluster HDFS à volonté en ajoutant de nouveaux modules sans pour autant avoir à revoir toute l'architecture. En plus de cela le framework offre une bonne gestion des pannes, par exemple lorsqu'un disque dur tombe en panne, ses données ne sont pas perdues, mais répliquées sur un autre nœud du réseau. Tous les modules d'Hadoop sont en fait conçus dans l'idée fondamentale que les pannes matérielles sont fréquentes et qu'en conséquence elles doivent être gérées automatiquement par le framework. Le framework prend même en charge les tâches de planification, de suivi et de réexécution des tâches échouées.

2.3. Framework Apache Spark

D'autres technologies visant à tendre vers des traitements plus « temps réel » ont émergé dans la foulée, c'est le cas d'Apache Spark. En effet, pour certains traitements qui se font de manière itérative, c'est à dire qui effectuent plusieurs tours de calcul sur les mêmes données, Hadoop ne convient pas parce qu'il manque de support intégré pour le processus itératif (Marone et al., 2019). En effet, Hadoop utilise le modèle de programmation **Map Reduce** qui lit et écrit à partir du disque, ce qui ralentit la vitesse de traitement. Ainsi, Apache Spark a été présentée comme une alternative à Hadoop pour surmonter ces problèmes (Marone et al., 2019). Apache **Spark** est un framework de calcul itératif

basé sur la mémoire, permettant de disposer d'un surcroît de puissance via l'utilisation de techniques in-memory de traitements de données permettant de stocker l'ensemble du processus en mémoire. Il réduit ainsi le nombre de cycles de lecture / écriture sur le disque. C'est pour cela qu'il convient aux applications utilisant des opérations itératives. Par exemple, Spark convient très bien aux algorithmes de machine learning, qui ont besoin d'effectuer plusieurs opérations sur les données. C'est un outil très adapté pour les campagnes de marketing en temps réel, les recommandations de produits en ligne et même la cyber sécurité. C'est donc un framework de calcul en mémoire parallèle, très rapide, spécialisé dans le calcul de grands ensembles de données distribuées. Les expériences ont montré que **Spark** surpasse de deux ordres de grandeur, les travaux classiques de **Map Reduce** en termes de vitesse (Marone et al., 2019).

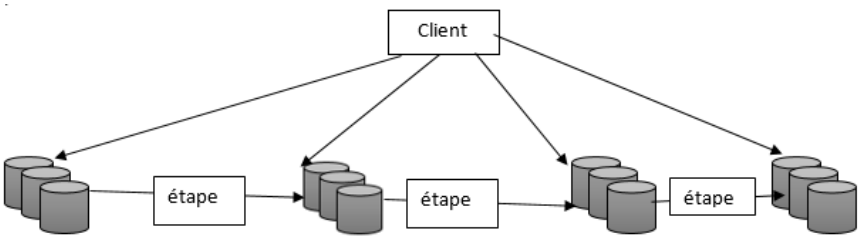

Figure 3 : Fonctionnement d'Hadoop

Comme nous le voyons à la figure précédente (figure 3), le traitement sur Hadoop se fait par étapes. La séquence de travail sur Hadoop ressemble à ceci : les données sont lus au niveau des disques durs du cluster, les opérations sont exécutés, ensuite les résultats sont enregistrés au niveau du cluster. Puis, on passe à l'étape suivante où une lecture est à nouveau faite au niveau des disques durs des machines sur le cluster et l'opération suivante est exécutée avec ces données mises à jour. Les résultats de cette opération sont à nouveau enregistrés sur les disques durs du cluster. etc.

Au contraire, Spark comme le montre la figure suivante (figure 4) peut travailler sur la totalité des données en une seule fois. Il exécute la totalité des opérations d'analyse de données en mémoire (RAM) et en temps quasi réel. Voici comment Spark procède : il lit les données au niveau des disques durs du cluster, effectue toutes les opérations d'analyses en autant d'étapes nécessaires en utilisant les RAM, écrit les résultats au niveau du cluster, et c'est tout ». Donc Spark ne fait pas des

aller-retour sur le disque dur, ce qui explique sa rapidité qui dépasse celle d'Hadoop.

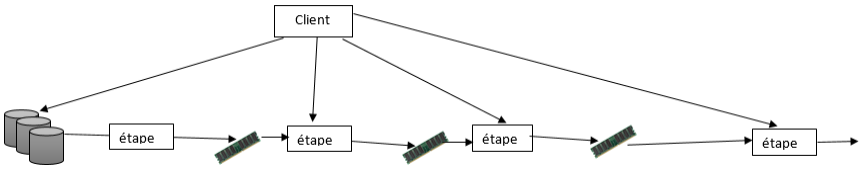

Figure 4 : Fonctionnement de Spark

Cependant, même si Spark sait traiter des données distribuées, il a besoin de s'appuyer sur un système de stockage distribué, car il ne sait pas faire du stockage distribué (Marone et al., 2019). Autrement dit Spark ne dispose pas de système de gestion de fichiers propre, ce qui signifie qu'il faut lui associer un système de fichiers. Étant donné que Spark a été conçu pour Hadoop (et qu'il fait partie de l'écosystème Hadoop), il est préférable de l'utiliser avec le système de fichiers HDFS d'Hadoop, avec lequel d'ailleurs il fonctionne mieux.

On peut alors constater que l'écosystème Hadoop est très riche et présente plusieurs avantages qui combinés au coût peu onéreux de stockage peut parfaitement convenir à l'archivage de grandes quantités de données. Hadoop est comme nous venons de le montrer, bien adapté à la réception et au traitement de données archivées. Il est d'ailleurs considéré par l'expert Big Data Dirk DEROSS de la société IBM (2014) comme une destination privilégiée pour les données à archiver. De plus l'intégration d'Hadoop avec les bases de données NoSQL rend le traitement et l'analyse des Big Data plus puissants. En effet, le paradigme Map Reduce, qui est le modèle conceptuel à la base d'Hadoop, utilise des bases de données distribuées non relationnelles afin de faciliter le traitement et surtout augmenter la vitesse d'exécution. Pour ce faire, il procède à la création de tables et d'index, non pas grâce aux techniques d'indexation des SGBD relationnels, mais avec des bases de données non relationnelles comme Hbase dont les tables peuvent servir d'entrée et de sortie pour des tâches Map Reduce. Cette combinaison des bases de données NoSQL avec Hadoop dépasse les solutions de GED ou d'archivage traditionnel qui présentent des limites dans leur approche Big data, car traiter les Big data demande des ressources que les solutions traditionnelles de gestion documentaire n'ont pour la plupart pas la capacité de fournir (Vieira, 2018).

On peut dire également que les Big data et les outils pour les gérer peuvent donner une nouvelle mission à l'archiviste : il s'agira désormais pour ce dernier, de jouer un rôle organisationnel, celui d'organiser, structurer au mieux les données en créant des index et des tables à partir de leur collection pour permettre l'analyse, les calculs sur les données et non plus seulement pour trouver les documents unitaires. L'archiviste doit donc comprendre les enjeux du Big data par rapport aux archives et permettre la valorisation de l'archive à l'ère du Big data afin que les organismes et les autres acteurs concernés puissent en tirer une valeur ajoutée.

3. Discussion

Le Big data a permis la mise au point de solutions complexes qui permettent de traiter et de mieux visualiser, analyser et cataloguer les grands volumes de données. Il fait évoluer l'archivage et pousse les systèmes de gestion de contenu documentaire à évoluer. On peut citer des solutions Big data comme Apache Hadoop/Spark et les bases de données NoSQL, qui sont des plateformes évolutives permettant de stocker et traiter les mégadonnées et qui peuvent ainsi aider les systèmes de gestion de contenu documentaire à répondre aux enjeux du Big data. Il est d'ailleurs probable que la plupart des utilisateurs choisiront pour leurs besoins, des logiciels spécifiques basés notamment sur les bases de données NoSql et l'écosystème Hadoop, et qui donnent de bons résultats en termes de performances et de scalabilité.

Toutefois il est à noter que, malgré tous les avantages que présente l'écosystème Hadoop, l'archivage dans Hadoop n'intègre pas nativement les fonctionnalités métiers qu'une SAE ou une GED peut posséder. Ce qui veut dire que des outils tels qu'Apache Hadoop/Spark et les bases de données NoSql peuvent être très performants techniquement parlant, mais pauvre fonctionnellement. Ils sont donc incapables de remplacer l'humain pour le moment, car celui-ci détient les compétences métiers et « fonctionnels ». Donc on peut dire qu'ils ne peuvent pas prendre la place de l'archiviste, car ils ne sauraient en aucun cas remplir toutes les fonctions qui sont attendues pour « l'archivage électronique » (Béchard et al., 2014). Dans Hadoop la notion d'archivage semble d'ailleurs, plus s'apparenter au stockage. Alors que l'archivage contrairement au stockage est une méthode visant à gérer et organiser l'information dans le temps, quel que soit son support, pour la rendre accessible durablement, bien au-delà de la durée de vie des supports (Ben Amor et Mkadmi, 2018). Ainsi archiver des mégadonnées n'est pas simplement réductible à l'utilisation de la technologie, les compétences de l'archiviste, son

approche et ses méthodes sont nécessaires réfutant ainsi tout déterminisme technologique.

A cela s'ajoute aussi, le fait que dans les Big data, sont présentes des données de toute sorte y compris des données personnelles et confidentielles. Mais de célèbres scandales ont montré la faible emprise qu'ont les utilisateurs sur leurs données personnelles et que certaines sociétés utilisaient la technologie pour développer à notre insu des stratégies marketing ou économique qui ont une réelle incidence sur le plan social et politique (Fazzini, 2018). C'est le cas du scandale Facebook-Cambridge Analytica appelée aussi la fuite de données Facebook-Cambridge Analytica dans lequel la société Cambridge Analytica (CA) est accusée d'avoir recueilli les données personnelles de plus de 87 millions d'utilisateurs Facebook (SOLON, 2018).

Cambridge Analytica est une entreprise britannique spécialisée dans le profilage psychographique, c'est à dire la segmentation ou le ciblage basé sur les styles de vies, les croyances, les valeurs, les personnalités des consommateurs (SOLON, 2018). Cette entreprise vend des services tels que la récupération de données sur les réseaux sociaux afin de tenter de prédire les comportements électoraux. Ce qui peut servir à des équipes d'organisations politiques et de campagnes électorales en leur permettant de peaufiner leur ciblage en vue du scrutin. Elle aide les partis politiques à cerner les portions de la population pour lesquelles il serait gagnant de déployer des efforts afin de les convaincre de voter pour eux.

Cambridge Analytica a d'ailleurs été connue pour son travail avec l'équipe de Donald Trump lors de l'élection présidentielle de 2016. Elle a aussi travaillé avec d'autres candidats républicains, comme Ted Cruz ou Ben Carson.

Cette société a été accusée d'avoir utilisé des données qui ont été prises sans l'autorisation des utilisateurs de Facebook violant ainsi les conditions d'utilisation du réseau social (SOLON, 2018). Tout est parti d'une application nommée : This is your digital life qui a été utilisée par le chercheur Aleksandr Kogan, de l'Université de Cambridge en 2015. Cette application consiste à faire effectuer un test à un utilisateur de Facebook en lui soumettant un questionnaire à remplir. Ensuite l'application analyse les données personnelles entrées dans le questionnaire et dresse un portrait psychologique de l'utilisateur. Cependant, elle ne se limitait pas à dresser un portrait psychologique, mais permettait également de collecter des données sur les amis Facebook des utilisateurs qui effectuaient le test. Ainsi, si seulement 270 000 personnes ont participé

au test, des données de plus de 50 millions de personnes ont également été collectées.

Le chercheur avait alors utilisé l'application et donné les informations qu'il a recueillies à Cambridge Analytica pour qu'elle les analyse. Cette dernière les a ensuite utilisées pour peaufiner son expertise en marketing politique. Cependant, les utilisateurs qui ont accepté de donner leurs informations personnelles au chercheur n'étaient pas au courant qu'elles seraient transmises à Cambridge Analytica pour qu'elles soient utilisées à des fins politiques. Au contraire, ils se sont fait dire que ces données étaient collectées à des fins scientifiques.

Ainsi un nouveau règlement sur la protection des données (RGPD) a été mis en place par la CNIL (Commission nationale de l'informatique et des libertés) et vise à protéger les données personnelles des utilisateurs. Dans ce règlement un droit à l'oubli numérique est prôné, la responsabilisation des entreprises quant à la sécurité et à la conservation des données à caractères personnelles, la mise en place d'audits et d'analyse de conformité ou encore la création d'un poste de DPO (Data Protection Officer) (Brasseur, 2016). Ce règlement devrait donc pousser les entreprises à être plus transparent sur la gestion des données utilisateurs, tout comme la neutralité de l'internet.

En ce sens, le spécialiste de l'information a son rôle à jouer. Par exemple l'archiviste devrait veiller à ce que certaines données soient conservées selon des normes de sécurité claires et conformes aux exigences de la loi, selon une durée de conservation bien définie. De cette façon les archives intermédiaires destinées au versement dans un fonds d'archive, devront respecter une politique de conformité à la loi, d'intégrité, de fidélité et ne devront pas être perdues. L'archiviste qui est à la fois record manager et « conservateur », possède les compétences requises et peut apporter son expertise de gestionnaire de l'information afin d'optimiser l'environnement Big data des institutions.

Conclusion

Les Big data (ou mégadonnées en français) rassemblent des données massives et variées (du texte, du son ou des images…) et qui n'ont a priori pas de lien entre elles. Ces données sont stockées sur une base numérique, puis traitées grâce à des technologies avancées adaptées qui permettent d'élaborer des diagnostics, de prendre des décisions en conséquence et d'établir des plans d'action. Ces données non structurées s'accumulent de plus en plus dans les organisations privées comme publiques et représentent dès lors un défi pour les spécialistes de

l'information tels que les archivistes, qui doivent les gérer pour en tirer parti (Ben Amor et Mkadmi, 2018). De la même manière, cette augmentation des volumes de données non structurées challengent les solutions de gestion documentaires classiques qui ont du mal à faire face à un tel volume et une telle variété (Phillipon et Besson, 2020).

Pour faire face à ces mégadonnées, un changement du socle technologique traditionnel fourni par la plupart des systèmes d'archivage traditionnel s'impose (Goupil, 2020). C'est pour cela qu'il est nécessaire de faire appel aux nouvelles technologies Big data ainsi qu'à des technologies émergentes telles que l'intelligence artificielle (IA/ML) et la Blockchain, afin de permettre aux organisations d'exploiter leurs données documentaires de manière plus efficace et à moindre coût. Ces dernières années, une tendance vers la technologie Big data a été d'ailleurs constatée, et comme le souligne Forrester, les technologies Big data ouvrent de nouvelles possibilités à l'archivage des données, en tirant parti des normes ouvertes, de l'intégration, de plates-formes évolutives telles que Hadoop et NoSQL qui peuvent stocker et traiter de très gros volumes de données structurées, non structurées et semi-structurées et permettent la recherche, la découverte et l'exploration à des fins de conformité et d'analyse (Yuhanna et al., 2015).

Cependant l'archivage dans les Big data n'est pas qu'une question de technologie, il est nécessaire d'y associer l'expertise de l'archiviste, ses méthodes ainsi que son approche. En effet ces technologies bien que performantes ne saurait prendre la place de l'humain pour le moment, elles ne peuvent pas se substituer à l'archiviste, pour remplir toutes les fonctions nécessaires à l'archivage électronique (Béchard et al., 2014).

L'un des rôles majeurs de l'archiviste devrait naturellement être de : permettre de déployer des méthodes d'archivage qui constitueront, une couche fonctionnelle qui sera intégrée à ce socle technique et permettra de mieux gérer le cycle de vie des données dans le temps. L'archiviste dans cet environnement technologique peut jouer un rôle au niveau de la gouvernance de l'information qui recouvre des aspects de confidentialité, d'intégrité et d'accessibilités. Il doit être en mesure d'agréger des données en provenance de différents outils appartenant au SI (Systèmes d'Information) de l'organisation, de mettre en place des plans de classification, d'identifier et supprimer les données redondantes, aider à reconnaître les documents sélectionnés qui sont restés intègres et permettre aux utilisateurs concernés de les exploiter plus facilement. L'archiviste a toutes les potentialités pour être celui qui veille à placer les données dans leur contexte pour qu'elles conservent leur sens métier et

aider ainsi à résoudre le problème du vrac informationnel, qui ne peut pas être résolue par l'utilisation seule de la « technologie ». L'archiviste devrait même être associé à la conception du SI dès le départ afin que les aspects de gouvernance documentaire soient pris en compte. Pour ce faire, il devra se doter de compétences numériques certaines et collaborer étroitement avec les services informatiques.

Cependant une question peut se poser, les étudiants formés dans les universités et institutions et qui sont censés devenir les futurs professionnels des archives, reçoivent-ils, au cours de leur formation, les nouvelles compétences et connaissances technologiques requises pour faire face à l'archivage des Big data ?

Bibliographie

- BASTIN Gilles ; TUBARO Paola. Presses de Sciences Po | « Revue française de sociologie » 2018/3 Vol. 59 | pages 375 à 394, ISSN 0035-2969, ISBN 9782724635669. Article disponible en ligne à l'adresse : https://www.cairn.info/revue-francaise-de-sociologie-2018-3-page-375.htm.

- BATTISTI Michèle, (2015). Big data et Smart culture. I2D, Information, données et documents 52, 2 (2015), 15–17. https://www.cairn.info/revue-i2d-information-donnees-et-documents-2015-2-page-15.htm.

- BÉCHARD Lorène ; FUENTES HASHIMOTO Lourdes ; VASSEUR Édouard, (2014). Les archives électroniques. Paris : Association des archivistes français. *Les Petits guides des archives*. ISBN 978-2-900175-06-4. CD974.4, page 10.

- BEN AMOR Fatma ; MKADMI Abderrazak, (2018). Les Archives à l'Ère des Big Data : Les Enjeux de l'Archivage des Données Numériques Massives. *In Digital Tools & Uses Congress* (DTUC » 18), October 3–5, 2018, Paris, France. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3240117.3240139.

- BRASSEUR Christophe, (2016). Enjeux et usages du Big Data. *2e édition. Paris : Hermès Lavoisier. Management et informatique*. ISBN 978-2-7462-4758-1. 658.05, page 78.

- CHARAUDEAU Marie-Odile ; FRITEL Alexis ; HUOT Charles ; MARTIN Philippe ; PRÉVEL Laurent. Et demain ? Archivage et big data. *La Gazette des archives* [en ligne]. 2015, no 240, p. 373‑384. Disponible à l'adresse : http://www.archivistes.org/Voyages-extraordinairement-numeriques-10-ans-d-archivage-electronique-et-demain.

- CHARLIN Laurent. Intelligence artificielle : une mine d'or pour les entreprises. Dans *Gestion 2017*/1 (Vol. 42), pages 76 à 79.

- DEROOS Dirk, (2014). Hadoop as an Archival Data Destination - dummies. [en ligne]. [Consulté le 26 décembre 2017]. Disponible à l'adresse : http://www.dummies.com/programming/big-data/hadoop/hadoop-as-an-archivaldata-destination/.

- ESPINASSE Bernard ; BELLOT Patrice, (2017). Introduction au Big Data — Opportunités, stockage et analyse des mégadonnées. *Techniques de l'ingénieur* [en ligne]. 10.02.2017. [Consulté le 30.10.2018]. Disponible à l'adresse : https://www.techniques-ingenieur.fr/base-documentaire/innovation-th10/les-technologies-numeriques-au-service-de-la-ville-et-de-la-personne-42674210/introduction-au-big-data-h6040/ [accès par abonnement].

- ESTEVA Mariam ; Jerey Felix Tang ; Weijia Xu ; Karthik Anantha Padmanabhan. 2013. DataMining for 'Big Archives' Analysis: a Case Study. https://onlinelibrary. wiley.com/doi/pdf/10.1002/meet.14505001076.

- GOUPIL Philippe. La Ged, c'est fini !. Dans *I2D – Information, données & documents2017*/2 (Volume 54), page 18.

- HAQUETTE Paul, consultant Eskimoz. Le big data libère le potentiel des archives d'entreprises. avril 2019. https://fr.blog.businessdecision.com/big-data-potentiel-archives-entreprises/.

- JOLLY Martyn, (2009). Big Archives and Small Collections: Remarks on the Archival Mode in Contemporary Australian Art and Visual Culture. *Public History Review21* (2009), 60–80.

- LAURENT Pascale ; LOWINGER Hélène ; MILLET Jacques ; CALDERAN Lisette, (2015). Big data: nouvelles partitions de l'information actes du séminaire IST Inria, octobre 2014. Louvain-la-Neuve [Paris] : De *Boeck ADBS*. ISBN 978-2-8041-8915-0. 004, page 33.

- LEMBERGER Pirmin ; BATTY Marc ; MOREL Médéric ; RAFFAËLLsI Jean-Luc ; GÉRON Aurélien, (2016). Big data et machine learning: les concepts et les outils de la data science. 2e éd. Paris : *Dunod. InfoPro*. ISBN 978-2-10-075463-2. Q325.5.

- MARONE R. M. N. ; CAMARA F. ; NDIAYE S. ; KANDE D., (2019) Two Parallelized Filter Methods for Feature Selection Based on Spark. In: Zitouni R., Agueh M. (eds) *Emerging Technologies for Developing Countries*. *AFRICATEK 2018*. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 260. Springer, Cham.

- MORDELL Devon. Critical Questions for Archives as (Big) Data. Published by *Association of Canadian Archivists*. *Archivaria*, 87 (Spring 2019), pp. 140-161 (Article).

- MOUGAYAR William, auteur de Le Business Blockchain. « L'identité et la réputation en ligne seront décentralisées. Nous serons les propriétaires de nos données. »: *Promesses, pratique et application de la prochaine technologie Internet (2016)*.

- NIEUWBOURG Philippe, (2017). Le concept de » data lake' — lac de données : explication de texte. https://www.decideo.fr/.Le-concept-de-data-lake-lac-de-donnees-explication-de-texte_a6976.html.

- PHILLIPON Eric ; BESSON Jérôme. QUEL FUTUR POUR LA GESTION ELECTRONIQUE DE DOCUMENT ? 2020 — Sentelis | Cabinet de conseil en stratégie, gouvernance et architecture des systèmes d'information. https://www.sentelis.com/articles/quel-futur-pour-la-gestion-electronique-de-document-1.

- SEGUY Pierre. Mémoire de fin d'études : *Exploitation des données personnelles Enjeux et perspectives du data marketing appliqué à la diffusion des films en salles*. Tuteur : Antoine Gouiffes-Yan, 21 avril 2019 Sous la direction d'Eric Vicente et Etienne Ollagnier.

- SOLON Olivia, « Facebook says Cambridge Analytica may have gained 37m more users' data » [archive], sur *the Guardian*, 4 avril 2018 (consulté le 6 avril 2021).

- VIEIRA Alexandre. Mémoire de fin d'étude / aout 2018 : *Archiver les Big Data : un enjeu pour l'archiviste d'aujourd'hui et de demain ?* Sous la direction de Laurent Duplouy Maitre de Conférence — ENSSIB. Université de LYON.

- YUHANNA Noel ; OWENS Leslie ; MCKINNON Cheryl ; CULLEN Elizabeth ; LYNCH Diane. In The Era Of Big Data, Archiving Is A No-Brainer Investment. For Enterprise Architecture Professionals. *Vendor Landscape: Big Data Archiving*. August 11, 2015.