Available online at https://www.asjp.cerist.dz/en/PresentationRevue/134

RIST

Information Processing at the Digital Age Journal

Volume 27. No. 1 (2023)

Introduction au BIG DATA : Concepts et Technologies

Faiza Deghmani[a,b]

[a]*Division Sciences de l'Information. Centre de Recherche sur l'Information Scientifique et Technique. Alger, Algérie.*
[b]*LSI, Université des Sciences et de la Technologie Houari Boumediene BP 32 El Alia 16111 Bab Azzouar Alger, Algérie.*

Résumé

Depuis quelques années, le terme Big Data s'est généralisé et les plus grandes entreprises et fournisseurs de données dans le monde y sont déjà passés. Ce phénomène qui a changé le monde, a vu le jour suite à l'explosion des données numériques et l'incapacité des systèmes traditionnels à gérer ces énormes quantités des données. En fait, Google, Yahoo et d'autres entreprises du web ont été les premiers confrontés aux problèmes de passage à l'échelle de leurs systèmes, ce qui a motivé le développement des premiers projets Big Data. Ainsi, pour répondre aux exigences des données de plus en plus massives, plusieurs projets ont été développés par la suite. Cet article est une introduction au Big Data et à ses technologies récentes.

Keywords : Big Data, Technologies du Big Data, Architecture Big Data, Modèles de données du Big Data;

1. Introduction

L'humanité produit aujourd'hui autant d'information en deux jours qu'elle ne l'a fait en deux millions d'années[*]. La généralisation de l'usage d'internet, l'accessibilité des coûts de stockage et la prolifération des réseaux sociaux constituent les raisons principales de l'émergence du Big Data. En effet, 5 exaoctets de données sont produits tous les deux jours ; c'est plus de données que les humains n'en ont généré depuis l'aube de la civilisation jusqu'en 2003. Ces données massives sont générées à une très grande vitesse, capturées et stockées pour diverses applications.

On parle depuis quelques années du phénomène Big Data, souvent traduit par « données massives ». Le nouveau terme « Big Data » est né du besoin des grandes entreprises et industries du web d'analyser de grandes quantités de données émergentes et qui nécessitaient de nouvelles technologies et architectures Yaqoob et al., 2016. En d'autres termes, le Big Data nécessite le nettoyage, le traitement, l'analyse et la

[*] https://fr.wikipedia.org/wiki/Big_data

protection des données ainsi que la fourniture d'un accès précis à des ensembles de données volumineux et évolutifs. Les entreprises réalisent de plus en plus que l'analyse des données est un facteur clé pour pouvoir rester compétitif sur le marché du travail Oussous et al., 2018.

En fait, les grands acteurs du web tel que Google et Yahoo étaient confrontés aux problèmes de scalabilité (passage à l'échelle) des systèmes et du temps de réponse aux requêtes utilisateurs. Alors pour faire face aux défis des données volumineuses, ils ont proposé les premiers projets Big Data. Très rapidement, Amazon et Facebook ainsi que d'autres entreprises ont suivi leur chemin et ont proposé leurs propres projets Big Data Chen et al., 2014.

Quelques années plus tard, le Big Data ne devient plus le centre d'attention des entreprises du web seulement ; de grandes nations s'intéressent aux projets Big Data en raison de la valeur intéressante qui peut être extraite de ces données massives. Les États-Unis ont été l'un des leaders à saisir l'opportunité du Big Data. En mars 2012, l'administration Obama a lancé le projet « Big Data Research and Development Initiative » avec un budget de 200 millions de dollar. Son objectif était d'améliorer la capacité du gouvernement américain à extraire des connaissances à partir de données numériques massives. Au Japon, le développement du Big Data est devenu un axe important de la stratégie technologique nationale en juillet 2012. Les Nations Unies ont publié un rapport intitulé « Big Data for Development: Opportunities and Challenges » qui met l'accent sur les principales préoccupations concernant les défis du Big Data et parle du comment le Big Data peut servir le développement international Oussous et al., 2018. Par conséquent, de nombreux modèles, des frameworks et de nouvelles technologies Big Data ont été créés pour fournir plus de capacité de stockage, de traitement parallèle et d'analyse en temps réel de différentes sources hétérogènes Oussous et al., 2018.

Dans ce contexte, le présent papier est une introduction aux différents concepts et technologies récentes développées pour le Big Data. Ce document est organisé comme suit. La section 2 est pour définir le Big Data et ses caractéristiques. La section 3 expose l'architecture du Big Data ; les différentes technologies de Big Data et les défis relevés sont présentés respectivement dans les sections 4 et 5. Et enfin une conclusion est donnée à la fin de l'article.

2. Définition du Big Data

Le terme Big Data fait référence à l'énorme quantité de données qui nécessite de nouvelles technologies et architectures pour extraire des informations précieuses à l'aide de méthodes analytiques nouvelles et innovantes. Divers éléments, notamment 3V (c'est-à-dire volume, variété et vitesse) et 4V (c'est-à-dire volume, vitesse, variété et véracité) ont été fournis pour définir le Big Data Oussous et al., 2018, Deghmani et al., 2021.

La plupart des data scientistes et experts dont Laney Doug Laney, 2001, ont décrit le Big Data à travers trois V nommés 3Vs, à savoir le volume, la vélocité et la variété Oussous et al., 2018, Yaqoob et al., 2016.

- Volume : Le terme volume fait référence à la quantité de données générées à partir de diverses sources Al-Mekhlal et Ali Khwaja, 2019, on estime qu'environ 2,5 exaoctets ont été générés chaque jour en 2012 Oussous et al., 2018. En 2013, le total des données numériques a été estimé par l'International Data Corporation[†] à 4,4 zettaoctets (ZB). Les données numériques sont passées à 8 ZB en 2015. Selon le rapport d'IDC, le volume de données atteint 40 zettaoctets en l'an 2020 Oussous et al., 2018.
- Variété : La variété décrit la diversité des sources et des types de données. Les données peuvent être structurées, semi-structurées et non structurées, ayant différents formats (texte, image, vidéo, son…etc.) et

[†] Une société qui publie des rapports de recherche

provenant de diverses sources telles que les sites web, les sites de médias sociaux, les e-mails et les documents Al-Mekhlal et Ali Khwaja, 2019.

- Vélocité : La vélocité concerne la vitesse des données entrantes et sortantes. Elle ne se limite pas seulement à la vitesse avec laquelle les données provenant de plusieurs sources sont injectées mais implique aussi la vitesse des flux de données tout au long du traitement des Big Data Al-Mekhlal et Ali Khwaja, 2019. Par exemple, Wallmart[‡] génère plus de 2,5 pétaoctets de données par heure à partir des transactions de ses clients. YouTube est un autre bon exemple qui illustre la rapidité du Big Data Oussous et al., 2018.

IBM et Microsoft ont rajouté la véracité ou la variabilité comme quatrième V pour définir le Big Data Yaqoob et al., 2016.

- Véracité : La véracité fait référence à la fiabilité et à la dimension qualitative des données. Traiter et gérer l'incertitude et les erreurs rencontrées dans certaines données représente un challenge de taille pour fiabiliser et minimiser les biais [§].

Par contre, McKinsey & Company[**] ont rajouté la valeur qui fait référence à la valeur des informations cachées dans le Big Data Yaqoob et al., 2016.

- Valeur : Cette caractéristique illustre la valeur que rajoute le Big Data ; les efforts et les investissements dans l'utilisation et l'application du Big Data n'ont de sens que si elles apportent de la valeur ajoutée Yaqoob et al., 2016.

3. Architecture du Big Data

Pour comprendre les fonctionnalités du Big Data, il faut d'abord comprendre son architecture et ses composants ; la figure 1 illustre l'architecture Big Data et ses composants clés qui devraient faire partie d'un système Big Data. Une architecture d'un système Big Data doit comporter les couches suivantes : sources de données, couche d'ingestion, couche de visualisation, couche de gestion de la plate-forme Hadoop, couche de stockage Hadoop, couche d'infrastructure Hadoop, couche de sécurité et couche de surveillance Erraissi et al., 2017.

[‡] Une chaîne internationale de magasins discount

[§] https://fr.wikipedia.org/wiki/Big_data

[**] Un cabinet international de conseil en stratégie

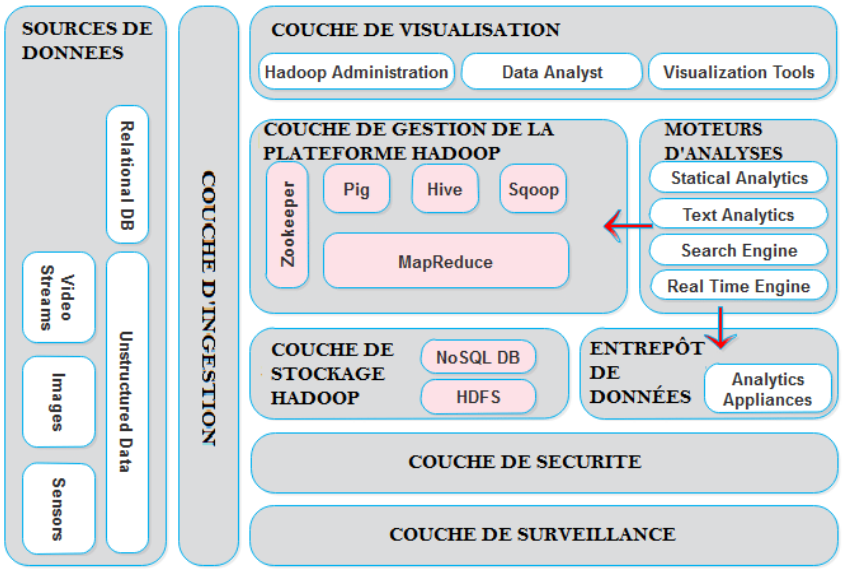

Fig. 1. Architecture du BIG DATA Erraissi et al., 2017

Source de données : Cette couche décrit les différents types de sources de données internes et externes qui doivent être analysées dans un système Big Data. Les données du Big Data se caractérisent par un volume, une variété, une vélocité et une valeur énormes. Il s'agit donc d'un flux de données complexe qui doit être parfaitement traité dans la couche d'ingestion Erraissi et al., 2017.

Couche d'ingestion : Cette couche permet de séparer le bruit des informations pertinentes. Elle doit être capable de gérer l'énorme volume, la grande vitesse et la variété des données. Elle doit également avoir la capacité de valider, nettoyer, transformer, réduire et intégrer les données afin que l'écosystème Hadoop puisse les utiliser ultérieurement Erraissi et al., 2017.

Couche de gestion de la plate-forme Hadoop : Cette couche fournit les outils nécessaires au traitement du MapReduce ainsi que les langages de requêtes afin d'accéder aux bases de données NoSQL en utilisant le système de fichiers de stockage distribué HDFS (PIG, HIVE, Sqoop, etc.) Erraissi et al., 2017.

Couche de stockage Hadoop : Cette couche est dédiée au stockage des données et diffère des anciennes technologies de stockage du fait que ce dernier soit massivement distribué. Le stockage Hadoop se base sur HDFS ; un système de fichiers distribué conçu pour stocker un très grand volume d'informations (téraoctets ou pétaoctets) à travers un grand nombre de machines dans un cluster. Il stocke les données de manière fiable, fonctionne sur du matériel de base, utilise des blocs pour stocker un fichier ou une partie de fichier, etc. Erraissi et al., 2017.

Couche de visualisation : Cette couche concerne l'aspect visuel et elle est utile pour les data analystes et les data scientistes pour mieux comprendre les données, et donc être de plus en plus en mesure d'examiner différents aspects des données dans différents modes visuels dans les plus brefs délais Erraissi et al., 2017.

Couche de surveillance : Avec autant de clusters de stockage de données distribués et de multiples points d'ingestion de sources de données, il est important d'obtenir une image complète du système Big Data grâce aux systèmes de surveillance. Par conséquent, cette couche définit les concepts utilisés par ces systèmes de surveillance pour augmenter les performances de Hadoop Erraissi et al., 2017.

Couche de sécurité : Du moment que la sécurité des données devient une préoccupation majeure dans tout système informatique, une couche de protection des données est conçue pour l'assurer. Les habitudes d'achat

des clients, les antécédents médicaux des patients, la démographie des maladies génétiques, tous ces types et utilisations de données et bien d'autres doivent être protégés, à la fois pour répondre aux exigences de conformité et pour protéger la vie privée de l'individu. Ces exigences de sécurité doivent faire partie de tout système Big Data dès le départ Erraissi et al., 2017.

4. Les modèles de données du Big Data

Les applications Big Data sont plus exigeantes en termes de concurrence, de latence, d'efficacité, d'économie de stockage, de conditions d'accès et de coûts opérationnels par rapport à ce qu'offrent les bases de données relationnelles. Ces Bases de données classiques ne sont pas conçues pour gérer des données massives et ne supportent pas les différentes caractéristiques des Big Data, ce qui a favorisé le développement d'un nouveau type de bases de données appelé NoSQL (Not only SQL) qui représente une alternative au relationnel mais ne le remplace pas. Google et Amazon étaient derrière le développement des premiers systèmes de gestion de bases de données (SGBD) NoSql Casado et Younas, 2015. En 2007, Amazon ainsi que d'autres entreprises, ont connu une énorme croissance de données et ont été confrontées au problème de gestion et de traitement de ces données. Pour réponde à ce besoin, Amazon a développé Dynamo, un SGBD NoSQL Casado et Younas, 2015. Un autre développement très important est venu de Google, qui a créé un nouveau magasin de données NoSQL appelé BigTable en 2008 Casado et Younas, 2015.

Plusieurs travaux montrent que les bases de données NoSQL offrent des avantages significatifs, tels qu'une mise à l'échelle facile et automatique, de meilleures performances et une haute disponibilité Deghmani et al., 2021. Une caractéristique clé des systèmes NoSQL est la mise à l'échelle horizontale sans partage, c'est-à-dire la réplication et le partitionnement des données sur de nombreux serveurs. Cela leur permet de prendre en charge un grand nombre d'opérations simples de lecture/écriture Casado et Younas, 2015.

Les systèmes NoSQL ne fournissent généralement pas de propriétés transactionnelles ACID[††] (atomicité, cohérence, isolation et durabilité) car ces dernières ne sont pas appliquées dans un contexte distribué, alors les mises à jour sont éventuellement propagées, mais avec des garanties limitées de cohérence des opérations de lecture Casado et Younas, 2015. Certains auteurs proposent le modèle BASE (Basically Available, Soft state, Eventually consistent) comme une alternative aux propriétés ACID Pritchett, 2008. En 2000, le professeur Eric Brewer a proposé le fameux théorème CAP (Consistency, Availability, and tolerance of network Partition); un concept clé pour comprendre les propriétés NoSQL Brewer, 2012. L'idée principale du théorème CAP est qu'un système distribué ne peut pas répondre simultanément aux trois besoins distincts mais seulement qu'à deux. Différents systèmes NoSQL ont été conçus dans le but d'atteindre les deux fonctionnalités spécifiées dans le théorème CAP Casado et Younas, 2015. Les systèmes fonctionnant sur une seule machine sont des exemples de systèmes CA - ils sont cohérents (car il n'y a pas de réplication) et disponibles. Les systèmes fonctionnant sur plusieurs machines sont des systèmes CP (MongoDB, HBase) ou des systèmes AP (Cassandra, CouchDB) Fraczek et Plechawska-Wojcik, 2017. La figure 1 classifie les différents systèmes NoSql existants selon le théorème de CAP.

[††] Un ensemble de propriétés qui garantissent qu'une transaction informatique est exécutée de façon fiable

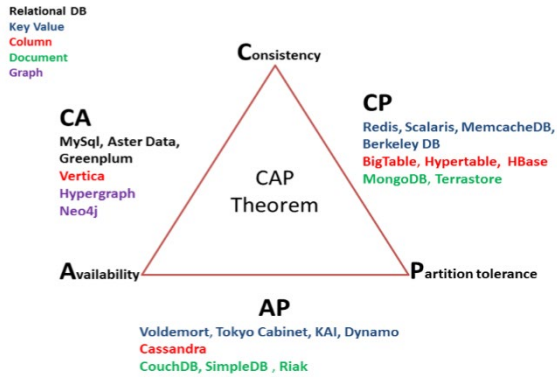

Fig. 1. Théorème de CAP[‡‡]

Les experts ont classé les bases de données NoSQL selon différents critères, notamment le modèle de données. Cependant, chaque type de bases de données NoSQL offre un certain niveau de flexibilité et un modèle de données différent pour répondre aux différents cas de Big Data. La classification des SGBD NoSql faite en fonction de leur structure de données ainsi que de leurs besoins de stockage et de restitution définit quatre classes : bases de données orientées document, clé-valeur, colonne et graphe Oussous et al., 2017.

Bases de données orientées Clé-valeur : ce type de bases de données permet au développeur d'applications de stocker des données sans schéma. Ces données consistent en une paire clé-valeur dont la clé est représentée par une chaîne, et les données réelles se sont la valeur. Les données peuvent être de n'importe quel type de données (une chaîne, un entier, un tableau, un objet…etc.). Ainsi, il assouplit l'exigence de données formatées pour le stockage, éliminant ainsi le besoin d'un modèle de données fixe Casado et Younas, 2015. Les fonctionnalités offertes par ces bases de données sont limitées aux valeurs de lecture, d'enregistrement et de suppression pour la clé spécifiée Fraczek et Plechawska-Wojcik, 2017. Des exemples de systèmes clé-valeur sont HBase, Cassandra et Redis Casado et Younas, 2015.

Bases de données orientées Document : Dans les bases de données orientées document (connu aussi sous le nom de magasin des documents) la structure de stockage sous-jacente utilisée est un document. Chaque magasin de documents diffère dans sa mise en œuvre des données. Le format de document est standard et peut être XML, JSON, BSON, PDF ou Microsoft Office. Chaque document est représenté par une clé unique, qui est une chaîne (URI ou chemin). Un langage de requête est fourni pour une récupération rapide des documents sur la base de leur contenu. MongoDB et CouchDB sont les SGBD les plus fameux Casado et Younas, 2015.

Bases de données orientées Colonne : Contrairement aux bases de données relationnelles, ces bases de données stockent leurs données dans des familles de colonnes organisées en lignes ; les lignes d'une même famille de colonnes peuvent avoir des colonnes différentes Fraczek et Plechawska-Wojcik, 2017. Les bases de données orientées colonne sont comparativement efficaces que les bases de données traditionnelles qui sont orientées lignes. Des exemples de systèmes orientés colonne sont Cassandra et Hypertable Casado et Younas, 2015.

[‡‡] https://openclassrooms.com/fr/courses/4462426-maitrisez-les-bases-de-donnees-nosql/4462471-maitrisez-le-theoreme-de-cap

Bases de données orientées graphe : Les bases de données graphe n'ont pas de schéma et sont basées sur un modèle mathématique « graphe ». Ils stockent les données dans les sommets du graphe et les relations entre les données dans les arêtes qui relient les sommets Fraczek et Plechawska-Wojcik, 2017. Les nœuds (ou sommets) peuvent représenter des entités telles que des personnes, des entreprises. Les propriétés désignent toute information relative aux nœuds. D'autre part, les arêtes reliant un nœud à un autre nœud, décrivent les liens entre les entités par exemple la relation d'amitiés entre des personnes. Des exemples de bases de données graphe sont Neo4J, Apache Giraph et Pregel de Google Casado et Younas, 2015.

5. Technologies du Big Data

Les principales technologies du Big Data peuvent être classées dans quatre catégories selon le rôle qu'elles fournissent[§§]. La figure 2 montre les différentes catégories. Nous détaillons dans cette section les outils les plus connus.

Fig. 1. Catégories de technologies Big Data[***]

5.1. Stockage de données

- Hadoop[†††] : Lorsqu'il s'agit de gérer le Big Data, Hadoop est l'une des principales technologies qui entrent en jeu. Hadoop est un framework libre et open source, utilisé pour le stockage et le traitement des Big Data. Son but est de faciliter la création d'applications distribuées (au niveau du stockage des données et de leur traitement) et échelonnables (scalables) permettant aux applications de travailler avec des milliers de nœuds et des pétaoctets de données. Ainsi chaque nœud est constitué de machines standard regroupées en grappe. Le noyau de Hadoop est constitué d'une partie de stockage HDFS (Hadoop Distributed File System), et d'une partie de traitement appelée MapReduce. Hadoop fractionne les fichiers en gros blocs et les distribue à travers les nœuds du cluster. Le framework Hadoop de base se compose des modules suivants : Hadoop Common ; le système de fichiers Hadoop Distributed File System (HDFS); Hadoop YARN et Hadoop MapReduce.

[§§] https://www.analytixlabs.co.in/blog/big-data-technologies/#Top_Big_Data_Technologies_Techniques

[***] https://www.javatpoint.com/big-data-technologies

[†††] https://fr.wikipedia.org/wiki/Hadoop

- MongoDB[‡‡‡]: MongoDB est un SGBD orienté document, qui fait partie des bases de données Nosql, répartissable sur un nombre quelconque d'ordinateurs et ne nécessitant pas de schéma prédéfini des données. La structure du stockage des données dans MongoDB est différente des bases de données RDBMS traditionnelles, il permet de manipuler des objets structurés au format BSON (JSON binaire), sans schéma prédéterminé. Les données prennent le format « document ».
- Cassandra [§§§] : Cassandra est un SGBD de type NoSQL conçu pour gérer des quantités massives de données sur un grand nombre de serveurs, assurant une haute disponibilité en éliminant le point de défaillance unique. Cassandra est basée sur une structuration en paires clé-valeur. L'architecture relationnelle est orientée colonne, avec des éléments plus traditionnels (stockage horizontal des paires). Des tables peuvent être créées, supprimées ou modifiées pendant l'exécution. Par contre à la différence des SGBD relationnel, Cassandra ne peut ni faire de jointures, ni sous-requêtes. Cassandra privilégie la dénormalisation des données.

5.2. Fouille des données

La fouille des données ou l'exploration de données a pour objectif l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données en utilisant des méthodes automatiques ou semi-automatiques. Plusieurs projets ont été proposés pour assurer la fouille de données dans une application Big Data, dont ElasticSearch et Apache Solr sont les plus connus.

- ElasticSearch[****]: ElasticSearch est un moteur de recherche et d'analyse utilisant la bibliothèque Lucene, et qui permet d'indexer, rechercher et analyser des données. Il opère en quasi temps réel et sur de grands volumes de données. Il fournit un moteur de recherche distribué et multi-entité à travers une interface REST. C'est la solution de recherche la plus populaire, elle est notamment utilisée par Netflix, Facebook, Deezer, Microsoft….
- Apache Solr : Solr est une plate-forme de recherche open source écrite en Java, basée sur la bibliothèque Lucene. Ses fonctionnalités sont la recherche en texte intégral, la mise en surbrillance des résultats, la recherche à facettes, l'indexation en temps réel, l'intégration de bases de données, les fonctionnalités NoSQL et les documents riches (par ex., Word, PDF) Aydoğan et al., 2016.

5.3. Analyse des données

L'analyse de données volumineuses décrit le processus de découverte de tendances, de modèles et de corrélations dans de grandes quantités de données brutes pour fournir un outil d'aide à la décision. Ce processus utilise des techniques d'analyse statistiques familières, telles que le clustering et la régression, et les applique à des ensembles de données plus étendus à l'aide d'outils plus récents[††††] ; Apache Kafka et Apache Spark sont parmi les outils d'analyses des Big Data.

- Apache Kafka: Apache Kafka est une plateforme de streaming populaire, connue pour ses trois fonctionnalités principales : éditeur, abonné et consommateur. Le projet Apache Kafka qui a été initialement développé par LinkedIn, vise à fournir un système unifié, en temps réel à latence faible pour la manipulation de flux de données. La plateforme qui est utilisée aussi par Netflix et Spotify est d'une plate-forme de streaming distribuée. Elle est également définie comme un système de courtage de messagerie directe et

---

[‡‡‡] https://fr.wikipedia.org/wiki/MongoDB

[§§§] https://fr.wikipedia.org/wiki/Cassandra_(base_de_donn%C3%A9es)

[****] https://www.3c-e.com/pourquoi-nous-utilisons-elasticsearch

[††††] https://www.tableau.com/learn/articles/big-data-analytics

asynchrone qui peut ingérer et effectuer un traitement sur des données de diffusion en temps réel. Cette plateforme est presque similaire à un système de messagerie d'entreprise ou à une file d'attente de messagerie[‡‡‡‡].

- Apache Spark : Apache Spark est l'une des technologies de base du Big Data. Apache Spark est un framework de calcul distribué, open source , connu pour ses capacités de calcul en mémoire qui contribuent à améliorer la vitesse globale du processus opérationnel. Il s'agit d'un ensemble d'outils et de composants logiciels structurés selon une architecture définie. Spark réalise une lecture des données au niveau du cluster (grappe de serveurs sur un réseau), effectue toutes les opérations d'analyse nécessaires, puis écrit les résultats à ce même niveau[§§§§].

5.4. Visualisation des données

Tableau est l'un des outils de visualisation de données les plus rapides et les plus puissants utilisés dans l'informatique décisionnelle. Il aide à analyser les données à une vitesse très rapide et aussi à créer des visualisations et des informations sous la forme de tableaux de bord et de feuilles de calcul.

6. Les défis du Big Data

Les chercheurs et les professionnels sont confrontés à plusieurs défis lorsqu'ils explorent des ensembles de mégas donnés et lorsqu'ils extraient de la valeur et des connaissances de ces mines d'information. Parmi les défis relevés par le Big Data, nous citons la capture, le stockage, la recherche, le partage, l'analyse, la gestion ainsi que la visualisation des données. De plus, il existe des problèmes de sécurité et de confidentialité, en particulier dans les applications distribuées pilotées par les données Oussous et al., 2018.

Dans cette section nous allons détailler quelques défis de Big Data qui sont encore ouverts à la recherche.

6.1. Gestion de données

Collecter, intégrer et stocker des données massives provenant des sources distribuées représente un défi très important pour les data scientistes ayant à faire à des Big Data. Il est crucial de pouvoir gérer efficacement ces énormes quantités de données afin de faciliter l'extraction d'informations fiables et d'optimiser les dépenses ce qui fait de la gestion du Big Data un défi major. En effet, une bonne gestion des données est la base de l'analyse du Big Data Oussous et al., 2018.

La gestion du Big Data signifie nettoyer les données pour la fiabilité, agréger les données provenant de différentes sources et coder les données pour la sécurité et la confidentialité. Cela signifie également assurer un stockage Big Data efficace et un accès basé sur les rôles à plusieurs terminaux distribués. En d'autres termes, l'objectif de la gestion du Big Data est de garantir des données fiables, facilement accessibles, gérables, correctement stockées et sécurisées Oussous et al., 2018.

6.2. Nettoyage des Big Data

Le nettoyage des données, l'agrégation, l'encodage, ainsi que le stockage et l'accès ne présentent pas de nouveaux défis mais existaient dans le cas de gestion de données traditionnelle. Le défi du Big Data est de

---

[‡‡‡‡] https://www.javatpoint.com/big-data-technologies

[§§§§] https://fr.wikipedia.org/wiki/Apache_Spark

savoir comment gérer la complexité de la nature du Big Data (vitesse, volume et variété) et la traiter dans un environnement distribué avec une variété d'applications Oussous et al., 2018.

En fait, pour des résultats d'analyse fiables, il est essentiel de vérifier la fiabilité des sources et la qualité des données avant d'engager des ressources. Cependant, les sources de données peuvent contenir des bruits, des erreurs ou des données incomplètes. Le défi est de savoir comment nettoyer des ensembles de données aussi volumineux et comment décider quelles données sont fiables et quelles données sont utiles Oussous et al., 2018.

6.3. Analyse des Big Data

Malgré que le Big Data offre de grandes opportunités et un potentiel de transformation pour divers secteurs, mais il présente également des défis sans précédents pour exploiter des volumes de données aussi importants et croissants. Ce qui fait qu'une analyse avancée des données est nécessaire pour comprendre les relations entre les fonctionnalités et explorer ces données. Par exemple, l'analyse des données permet à une organisation d'extraire des informations précieuses et de surveiller les modèles qui peuvent affecter positivement ou négativement l'entreprise Oussous et al., 2018.

D'un autre côté, les applications basées sur les données comme la navigation, les réseaux sociaux, la finance, la biomédecine nécessitent également une analyse en temps réel. Alors, des algorithmes avancés et des méthodes efficaces d'exploration de données sont nécessaires pour obtenir des résultats précis, pour surveiller les changements dans divers domaines et pour prédire les observations futures. De nos jours, il existe diverses techniques d'analyse, notamment l'exploration de données, la visualisation, l'analyse statistique et l'apprentissage automatique qui tentent de faire face à ce défi Oussous et al., 2018.

7. Conclusion

Le phénomène Big Data a émergé avec l'explosion de volume des données due à l'augmentation de la production de données numériques ces vingt dernières années et l'incapacité des systèmes traditionnels à gérer ces données. En effet, le terme Big Data fait référence aux grandes quantités de données qui nécessitent de nouvelles technologies et architectures pour extraire des informations précieuses. Pour répondre aux exigences du Big Data, plusieurs projets ont été proposés par les grands acteurs du web tel que Google, Amazon et Facebook. Le Big Data représente un écosystème technologique très large ; de nouveaux modèles de données pour le stockage, des outils d'exploration des données basés sur des méthodes automatiques ou semi-automatiques, outils d'analyse et de visualisation des données.

Cet article se voulait une introduction aux Big Data. Nous avons défini le Big Data en combinant plusieurs définitions existantes, ensuite nous avons présenté l'architecture du Big Data en soulignant ses couches les plus importantes. Les bases de données traditionnelles ne répondent plus aux exigences du Big Data, donc ce dernier requiert un nouveau type des bases de données nommées NoSql qui ont été abordées dans cet article. En outre, Nous avons présenté les principales technologies du Big Data et classé selon le rôle qu'elles fournissent. Nous avons, enfin, évoqué quelques défis que confrontent les chercheurs dans le Big Data et qui sont encore ouverts à la recherche.

Bibliographie

Yaqoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., Vasilakos, A. V., 2016. Big data : From beginning to future. *International Journal of Information Management*, *36*(6), 1231‑1247. https://doi.org/10.1016/j.ijinfomgt.2016.07.009

Oussous, A., Benjelloun, F.-Z., Ait Lahcen, A., Belfkih, S., 2018. Big Data technologies : A survey. *Journal of King Saud University* -

Computer and Information Sciences, *30*(4), 431‑448. https://doi.org/10.1016/j.jksuci.2017.06.001

Chen, M., Mao, S., Liu, Y., 2014. Big Data : A Survey. *Mobile Networks and Applications*, *19*(2), 171‑209. https://doi.org/10.1007/s11036-013-0489-0

Deghmani, F., AmineAmarouche, I., Boukhalfa, K., 2021. Applications of Graph Databases and Big Data Technologies in Healthcare. *Revue de l'Information Scientifique et Technique*, *26*(1), 47‑55.

Laney, D., 2001. 3D Data Management : Controlling Data Volume, Velocity, and Variety. *META group research note*, *6*(70), 1.

Al-Mekhlal, M., Ali Khwaja, A., 2019. A Synthesis of Big Data Definition and Characteristics. *2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, 314‑322. https://doi.org/10.1109/CSE/EUC.2019.00067

Erraissi, A., Belangour, A., Tragha, A., 2017. A Big Data Hadoop building blocks comparative study. *International Journal of Computer Trends and Technology*, *48*(1), 36‑40. https://doi.org/10.14445/22312803/IJCTT-V48P109

Casado, R., Younas, M., 2015. Emerging trends and technologies in big data processing. *Concurrency and Computation: Practice and Experience*, *27*(8), 2078‑2091. https://doi.org/10.1002/cpe.3398

PRITCHETT, D., 2008. BASE: An Acid Alternative : In partitioned databases, trading some consistency for availability can lead to dramatic improvements in scalability. *Queue*, *6*(3), 48‑55.

Brewer, E., 2012. CAP twelve years later : How the « rules » have changed. *Computer*, *45*(2), 23‑29. https://doi.org/10.1109/MC.2012.37

Fraczek, K., Plechawska-Wojcik, M., 2017. Comparative Analysis of Relational and Non-relational Databases in the Context of Performance in Web Applications. In *Beyond Databases, Architectures and Structures. Towards Efficient Solutions for Data Analysis and Knowledge Representation* (Vol. 716, p. 153‑164). Springer International Publishing. https://doi.org/10.1007/978-3-319-58274-0_13

Oussous, A., Benjelloun, F.-Z., Lahcen, A. A., Belfkih, S., 2017. NoSQL databases for big data. *Int. J. Big Data Intelligence*, *4*(3), 171-185.

Aydoğan, T., İlkuçar, M., Akca, M. A., 2016. An Analysis on the Comparison of the Performance and Configuration Features of Big Data Tools Solr and Elasticsearch. *International Journal of Intelligent Systems and Applications in Engineering*, *4*(Special Issue-1), 8‑12. https://doi.org/10.18201/ijisae.271328