

1. Introduction :

L'explosion quantitative des données numériques a obligé les chercheurs à trouver de nouvelles manières pour voir et d'analyser le monde. Il s'agit de découvrir de nouveaux ordres de grandeur concernant la capture, la recherche, le partage, le stockage, l'analyse et la présentation des données. Ainsi est née le « Big Data ». Il s'agit d'un concept permettant de stocker un nombre indicible d'informations sur une base numérique. [13]

Dans ce chapitre nous allons présenter le paradigme Big Data, ses caractéristiques ainsi que les Framework spécialisés dans le traitement de données massives.

2. Historique :

2.1. Le Big Data et l'histoire du stockage des informations :

Des premières formes d'écriture aux centres de données actuels, l'homme a toujours collecté des informations. L'avènement de la technologie a fait surgir un torrent de données qui nécessite des systèmes de stockage sophistiqués. Démarrant dans les années 1930, cette chronologie couvre le boom de l'information et illustre la manière dont le Big Data a initié le besoin d'organiser et de stocker les informations. [24]

2.2. Les débuts de la surabondance d'informations :

Il a fallu 8 ans pour publier les données du recensement effectué en 1880 aux États-Unis. Et il a été estimé que cette tâche aurait pris plus de 10 ans pour le recensement de 1890 avec les procédés de l'époque. Sans une nouvelle méthodologie, cette publication n'aurait donc pas été effectuée avant le recensement de 1900. [24]

2.3. La « Tabulating Machine » de Hermann Hollerith :

L'afflux de données du recensement amena l'invention par Hermann Hollerith de la machine mécanographique « Tabulating Machine » qui ordonnait les données et qui a permis d'effectuer la publication en environ 1 an. Hermann Hollerith devenait ainsi entrepreneur et sa société allait bientôt faire partie de la grande aventure que nous connaissons sous le nom d'IBM. [24]

2.4. L'essor de la population :

La surabondance d'informations s'est poursuivie avec la forte croissance de la population aux États-Unis, l'émission de numéros de sécurité sociale et l'accroissement global des connaissances (recherche), qui ont exigé une consignation des données plus structurée et plus complète. [24]

2.5. Les répercussions sur les bibliothèques :

Les bibliothèques, qui sont les premiers systèmes d'organisation et de stockage des données, ont dû adapter leurs méthodes de stockage afin de répondre à la croissance rapide des demandes de nouvelles publications et de nouvelles recherches. [24]

2.6. L'explosion de l'information :

Les spécialistes ont commencé à évoquer cette incroyable croissance de l'information sous le nom d'« explosion de l'information ». Utilisée pour la première fois en 1941 dans le journal « The Lawton Constitution », cette expression apparaît dans un article du magazine « New Statesman » en mars 1964, qui décrivait la difficulté de gérer les volumes d'informations disponibles. [24]

2.7. Première sonnette d'alarme en matière de stockage et de récupération des données :

Le premier avertissement que cette croissance des connaissances poserait un problème de stockage et d'accès aux données est apparu dès 1944. Cette année-là « Fremont Rider » bibliothécaire de la « Wesleyan University », a estimé que la taille des bibliothèques universitaires américaines doublait tous les seize ans. Compte tenu de ce taux de croissance, « Fremont Rider » a calculé que la bibliothèque de « Yale » comporterait en 2040 « environ 200 000 000 d'ouvrages, ce qui représenterait 10 000 kilomètres de rayons [...] [nécessitant] plus de 6 000 personnes pour leur référencement ». [24]

2.8. La théorie de l'information de Claude Shannon :

« Claude Shannon » a publié l'article intitulé « A Mathematical Theory of Communication » qui définit les critères minimums pour transmettre des informations à travers des canaux perturbés (imparfaits). Ce travail de référence a permis d'aboutir à la plupart des infrastructures actuelles. Sans cette nouvelle théorie, les données seraient bien plus volumineuses. L'étude de « Claude Shannon » fait suite à l'article « Certain Factors Affecting Telegraph Speed » de « Harry Nyquist », qui semble de prime abord assez éloigné, mais qui a permis d'échantillonner les signaux analogiques et de les représenter sous forme numérique afin de jeter les bases du traitement moderne des données. [24]

2.9. Mémoire virtuelle :

La notion de mémoire virtuelle a été développée par le physicien allemand « Fritz-Rudolf Güntsch », qui a étudié le stockage fini comme un concept infini. Le stockage géré par des logiciels et des matériels intégrés nous a permis de traiter les informations indépendamment des contraintes matérielles qui obligeaient précédemment à segmenter le problème. En effet, la solution consistait simplement à refléter l'architecture matérielle, bien que cela ne paraisse pas du tout naturel. [24]

2.10. Quand les connaissances scientifiques s'étendent :

« Derek Price » spécialiste de l'information, a généralisé les conclusions de « Fremont Rider » afin d'inclure pratiquement toutes les connaissances scientifiques. La révolution scientifique, comme il la nommait, était responsable de la transmission accélérée des nouvelles idées en tant qu'informations scientifiques. Cette croissance rapide s'est traduite par le doublement de la taille des journaux tous les 15 ans. [24]

2.11. À la recherche d'une solution organisationnelle :

Au début des années 1960, « Derek Price » a constaté qu'il était impossible de suivre le rythme de la recherche scientifique. Les résumés de journaux ont été créés à la fin du XIXe siècle afin de gérer la croissance des bases de connaissances. Ils enregistraient déjà une tendance similaire (multiplication par 10 tous les 50 ans) et avaient atteint une « ampleur critique ». Ce n'était donc plus une solution viable pour le stockage ou l'organisation des informations. [24]

2.12. Quand les systèmes informatiques centralisés entrent en scène :

Le secteur scientifique n'était pas le seul à être confronté à l'explosion des informations, c'était aussi le cas des entreprises. Avec l'afflux d'informations observé dans les années 1960, la plupart des entreprises ont commencé à concevoir, développer et mettre en œuvre des systèmes informatiques centralisés afin d'automatiser leurs systèmes d'inventaire. [24]

2.13. Base de données relationnelle :

En 1970 « Edgar F. Codd » mathématicien diplômé d'Oxford et travaillant dans le laboratoire de recherche d'IBM, a publié un article expliquant comment accéder à des informations stockées dans de grandes bases de données sans connaître la structure ni l'emplacement de ces informations. Jusqu'alors, la récupération des informations nécessitait des connaissances informatiques assez pointues, voire le recours aux services de spécialistes. Cette opération était à la fois chronophage et onéreuse. Aujourd'hui, la plupart des transactions de données de routine (accès aux comptes bancaires, utilisation de cartes de crédit, transactions boursières, réservations de voyage, achats en ligne) emploient des structures reposant sur la théorie de la base de données relationnelle. [24]

2.14. La montée de la communication bilatérale :

Le recensement des flux d'informations mené par le ministère des Postes et des Télécommunications au Japon a permis d'engager le suivi du volume d'informations en circulation dans le pays pour l'année 1975. S'appuyant sur le nombre de mots comme unité de mesure pour l'ensemble des médias, l'étude a démontré que l'offre d'information avait largement dépassé la consommation d'information, et que la demande de communication

unilatérale stagnait. Parallèlement, la demande de communication personnalisée bilatérale enregistrait une hausse, en réponse aux besoins des particuliers. [24]

2.15. Systèmes de planification des besoins en matériel (MRP) :

Au milieu des années 1970, des systèmes de planification des besoins en matériel (MRP) ont été conçus pour aider les entreprises manufacturières à organiser et planifier leurs informations. À cette même époque, les PC commençaient tout juste à intéresser les entreprises. Les processus métier et les fonctionnalités de comptabilité étaient alors au cœur des préoccupations, tandis que des entreprises comme Oracle, JD Edwards et SAP faisaient leur entrée dans le secteur. Oracle a finalement présenté le langage SQL (Structured Query Language) original et l'a commercialisé. [24]

2.16. Loi de Parkinson sur les données :

Alors que les informations commençaient à se développer plus rapidement, les possibilités de condenser le stockage et l'organisation des données se sont amoindries. Dans son discours intitulé « Where Do We Go From Here? », I.A. « Tjomsland » a déclaré : « Ceux qui connaissent les périphériques de stockage se sont rendu compte depuis longtemps que la première loi de Parkinson pouvait s'appliquer à notre secteur : « Les données s'étendent jusqu'à remplir l'espace disponible pour leur stockage ». Je pense que de grandes quantités de données sont conservées, car les utilisateurs n'ont aucun moyen d'identifier les données obsolètes. Les inconvénients du stockage des données obsolètes sont moins visibles que les inconvénients de la suppression de données potentiellement utiles. » [24]

2.17. Croissance des informations et secteur de la diffusion :

Alors que les technologies continuaient à se développer, chaque secteur a commencé à enregistrer des améliorations notables grâce à la mise en œuvre de nouvelles méthodes d'organisation, de stockage et de production des données. Puis les entreprises ont commencé à exploiter les données dans le but d'optimiser les décisions métier. Dans l'article Tracking the Flow of Information publié dans le magazine Science, « Ithiel de Sola Pool » a étudié la croissance des informations sur 17 supports de communication clés entre 1960 et 1977. Selon lui, la forte croissance des informations est due au développement du secteur de la diffusion. [24]

2.18. Systèmes de planification des ressources de production (MRP II) :

Après l'avènement des systèmes MRP, la planification des ressources de production (MRP II) a fait son apparition dans les années 1980, notamment en vue d'optimiser les processus de fabrication par la synchronisation des ressources et des exigences de production. La technologie MRP II incluait différents domaines comme la gestion de l'atelier et de la

distribution, la gestion de projet, la finance, les ressources humaines et l'ingénierie. Peu de temps après l'adoption de cette technologie, les autres secteurs (administrations, entreprises du secteur tertiaire, etc...) ont commencé à remarquer la technologie ERP, pour finalement l'adopter. [24]

2.19. Un besoin de fiabilité des données :

En 1985 « Barry Devlin » et « Paul Murphy » ont imaginé une toute nouvelle architecture de génération de rapports et d'analyse métier chez IBM (Devlin & Murphy, IBM Systems Journal 1988), qui est devenue par la suite la base de l'entrepôt de données (data warehousing). Cette architecture ou plutôt l'entreposage des données en général répond à un besoin de stockage cohérent et de grande qualité pour des données historiquement complètes et fiables. [24]

2.20. Une nouvelle dimension pour les systèmes logiciels :

Entre la fin des années 1980 et le début des années 1990, les systèmes de Progiciels de Gestion Intégrée (PGI) ou (ERP) ont enregistré un franc succès : plus sophistiqués, ils permettaient une coordination et une intégration dans toute l'entreprise. Les technologies de base des systèmes MRP, MRP II et ERP ont commencé à intégrer divers domaines, comme la fabrication, la distribution, la comptabilité, la finance, la gestion des ressources humaines, la gestion de projet, la gestion des stocks, l'entretien et la maintenance, ou encore le transport, offrant ainsi accessibilité, visibilité et cohérence à l'échelle de l'entreprise. [24]

2.21. Business Intelligence :

En 1989 « Howard Dresner » a développé la notion de « Business Intelligence (BI) », terme générique populaire inventé par « Hans Peter Luhn » en 1958. Selon « Howard Dresner » la Business Intelligence (ou l'analyse décisionnelle) désigne « des concepts et méthodes permettant d'améliorer la prise de décision métier grâce à des systèmes reposant sur des faits ». Peu après, en réponse au besoin d'une meilleure BI, des entreprises telles que (Business Objects, Actuate, Crystal Reports et MicroStrategy) ont vu le jour et commencé à proposer des rapports et des analyses de données d'entreprise. [24]

2.22. Le premier rapport sur base de données :

En 1992 « Crystal Reports » a créé le premier outil de reporting simple sous Windows. Ces rapports permettaient aux entreprises de créer un rapport unique des sources d'information multiples avec un minimum de code. Cela a permis de rendre supportable un environnement saturé de données et de rendre l'usage de la Business Intelligence accessible. [24]

2.23. L'explosion du World Wide Web :

Dans les années 1990, les nouvelles technologies ont littéralement explosé et les données de Business Intelligence ont commencé à s'accumuler sous la forme de documents Microsoft Excel. [24]

2.24. Une croissance phénoménale de la puissance informatique et d'Internet :

L'essor des données a occasionné d'autres défis pour les fournisseurs d'ERP. La nécessité de repenser les produits ERP, notamment dans le but de surmonter les obstacles inhérents aux droits de propriété et à la personnalisation, a obligé les fournisseurs à choisir une collaboration totalement transparente sur intranet. [24]

2.25. Le problème du Big Data :

Le terme « Big Data » a fait son apparition dans un article publié par « Michael Cox et David Ellsworth » chercheurs à la « NASA ». Tous deux affirmaient que l'augmentation du volume des données devenait problématique pour les systèmes informatiques de l'époque. C'est ce que l'on a appelé le « problème du Big Data ». [24]

2.26. L'avenir du stockage des données :

Dans son article intitulé (**How much information is there in the world ?**) « Michael Lesk » déclare « on compte peut-être quelques milliers de pétaoctets d'informations en tout et pour tout, et la production de bandes et de disques aura atteint ce niveau en l'an 2000. Donc dans quelques années, (a) nous serons en mesure de tout enregistrer (sans suppression d'informations) ; et (b) la plupart des informations ne seront jamais examinées par un être humain ». [24]

2.27. Quand les informations sont quantifiées :

« Peter Lyman » et « Hal R. Varian » de « l'UC Berkeley » a publié la première étude qui quantifiait, en termes de stockage informatique, le volume total d'informations initiales et nouvelles créées chaque année dans le monde. Cette étude, intitulée (**How Much Information ?**) A été établie en 1999, année durant laquelle le monde a produit environ 1,5 exaoctet d'informations. [24]

2.28. Les 3 V :

« Doug Laney » analyste chez « Gartner », a publié un rapport de recherche intitulé 3D Data Management : Controlling Data Volume, Velocity, and Variety.

Encore aujourd'hui, les 3 V constituent les critères globaux du Big Data. [24]

2.29. Services Web et ERP :

Les principaux fournisseurs d'ERP, comme SAP, PeopleSoft, Oracle et JD Edwards, commencent à s'intéresser de très près aux services Web afin de connecter leurs propres suites

logicielles, mais aussi pour permettre aux clients de créer facilement des applications à partir de données issues de plusieurs applications à l'aide du langage XML. [24]

2.30. Pleins feux sur la gestion des bases de données :

« Tim O'Reilly » a publié l'article intitulé (What is Web 2.0 ?), dans lequel il affirme que « les données sont le prochain Intel Inside ».

Toujours selon « Tim O'Reilly » : « Comme Hal Varian l'a expliqué au cours d'une conversation personnelle l'an dernier, "SQL est le nouvel HTML". La gestion des bases de données est une compétence essentielle pour les entreprises Web 2.0, tant et si bien que ces applications sont parfois appelées "infoware" et non plus simplement des "logiciels" ». [24]

2.31. Une solution ouverte face à l'explosion du Big Data :

Hadoop a été créé en 2006 pour que de nouveaux systèmes sachent gérer l'explosion des données Web. Pouvant être téléchargé, utilisé et amélioré gratuitement.

Hadoop constitue un moyen entièrement ouvert de stockage et de traitement des données qui « permet de traiter de manière parallèle et distribuée de grands volumes de données sur des serveurs standard et économiques, capables à la fois de stocker et de traiter les données, et d'évoluer à l'infini ». [24]

2.32. La prolifération des données se poursuit :

« Bret Swanson » et « George Gilder » ont estimé que le trafic IP aux États-Unis pourrait atteindre un zettaoctet d'ici à 2015 et que l'Internet américain de 2015 serait au moins 50 fois plus vaste qu'en 2006. [24]

2.33. Face au déluge de données, la méthode scientifique est obsolète :

Le terme « Big Data » a gagné en popularité dans les secteurs technologiques. Le magazine « Wired » a publié un article présentant les répercussions positives et négatives du « déluge de données » moderne. Dans cet article, « Wired » a annoncé « les débuts de l'âge du pétaoctet ». Bien que cette hypothèse soit tout à fait sensée, le terme « pétaoctet » s'est avéré trop technique pour le plus grand nombre. Inévitablement, ces pétaoctets, qui équivalent à des milliers de billions d'octets de données, annoncent des volumes de données encore plus importants : exaoctets, zettaoctets et yottaoctets. [24]

2.34. Des percées révolutionnaires :

Un groupe de chercheurs informatiques a publié un article intitulé par (Big Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science, and Society), dans lequel on peut lire : « À l'image des moteurs de recherche qui ont transformé notre manière d'accéder aux informations, d'autres formes d'informatiques du Big Data sont en mesure et en passe de transformer les activités des entreprises, des chercheurs, des médecins et des équipes de

renseignements et de défense gouvernementales... Il s'agit peut-être de la plus grande innovation informatique de ces dix dernières années. [24]

Nous commençons à peine à apercevoir son potentiel pour collecter, organiser et traiter les données dans tous les aspects du quotidien. Un petit investissement de la part du gouvernement (américain) pourrait grandement accélérer son développement et son déploiement. Cet aval a finalement conféré une certaine crédibilité intellectuelle au Big Data. [24]

2.35. Combien d'informations ?

Étude (How Much Information ? 2009 Report on American Consumers) du Global Information Industry Center révèle qu'en 2008, « les Américains ont consommé environ 1,3 billion d'heures d'informations, soit en moyenne près de 12 heures par jour. Cette consommation a atteint un total de 3,6 zettaoctets et de 10 845 billions de mots, ce qui correspond à 100 500 mots et 34 gigaoctets en moyenne par jour et par personne ». [24]

Un rapport de janvier 2011, intitulé (How Much Information ? 2010 Report on Enterprise Server Information), a ensuite estimé qu'en 2008, « les serveurs du monde entier ont traité 9,57 zettaoctets d'informations, soit pratiquement 1022 gigaoctets (soit dix millions de millions). Cela représentait 12 gigaoctets d'informations quotidiennes pour un employé moyen, ou environ 3 téraoctets d'informations par employé et par an. Les entreprises mondiales ont traité en moyenne 63 téraoctets d'informations par an ». [24]

2.36. Les données sont partout :

« The Economist » a publié un rapport intitulé (Data Everywhere), dans lequel « Kenneth Cukier » son auteur, écrit : « [...] le monde compte une quantité inimaginable d'informations numériques, qui croît très rapidement. [...] Tout le monde est touché, des entreprises aux secteurs scientifiques, et des gouvernements aux domaines artistiques ». [24]

2.37. La croissance réelle des données -2011- :

Un article intitulé (The World's Technological Capacity to Store, Communicate, and Compute Information) et tiré du magazine « Science » a estimé que la capacité mondiale de stockage des informations avait augmenté de 25 % par an entre 1987 et 2007.

Toujours d'après cette source 99,2 % du stockage des données était analogique en 1986, alors qu'il était à 94 % numérique en 2007. On a donc assisté à un bouleversement complet en à peine 20 ans (en 2002, le stockage numérique a surpassé le stockage non numérique pour la première fois). [24]

2.38. Capacité des informations :

En 2012, l'article (Tracking the Flow of Information into the Home), publié dans « l'International Journal of Communication », a calculé que la fourniture totale de médias dans les foyers des États-Unis était passée d'environ 50 000 minutes par jour en 1960 à près de 900 000 minutes par jour en 2005.

Toujours d'après cette source, les États-Unis « approchaient d'un millier de minutes de contenu média disponible par minute disponible pour la consommation ». [24]

2.39. Des questions critiques pour le Big Data :

L'article intitulé (Critical Questions for Big Data) et publié dans « l'Information, Communications, and Society Journal », définit le Big Data comme « un phénomène culturel, technologique et scientifique qui résulte de l'interaction entre :

- 1 La technologie (optimisation de la puissance de calcul et de la précision algorithmique pour rassembler, analyser, lier et comparer de grands ensembles de données). [24]
- 2 L'analyse (utilisation de grands ensembles de données pour identifier des schémas récurrents afin d'émettre des revendications économiques, sociales, techniques et juridiques). [24]
- 3 La mythologie (large croyance selon laquelle de grands ensembles de données promettent une forme supérieure d'intelligence et de connaissance, susceptible d'offrir un tout nouveau regard empreint de vérité, d'objectivité et de précision).» [24]

2.40. L'avenir du Big Data :

La production de données augmente à un rythme effréné. Les spécialistes penchent aujourd'hui pour une hausse de 4 300 % de la génération annuelle des données d'ici à 2020. Les facteurs de croissance incluent la transition des technologies analogiques vers le numérique, ainsi que la hausse rapide de la génération des données par les entreprises comme par les particuliers. [24]

3. Les Big Data :

3.1. Définition :

Les Big data sont un ensemble de solutions alternatives aux solutions traditionnelles de bases de données et d'analyse afin de traiter un volume très important de données, en temps réel et avec une très grande diversité de sources et de formats.

Les objectifs de ces solutions d'intégration et de traitements des données sont de traiter un volume très important de données aussi bien structurées que non structurées, se trouvant sur des terminaux variés (PC, smartphones, tablettes, objets communicants...), produites ou non en temps réel depuis n'importe quelle zone géographique dans le monde. [6] [15]

3.2.Caractéristiques des Big Data :

Pour caractériser les Big data on peut utiliser trois critères principaux. Appelés les 3V (Le volume, la Vitesse et la variété). [6]

- **Vélocité** : la vitesse à laquelle les données sont traitées simultanément. [6]
- **Variété** : l'origine variée des sources de données qui arrivent non structurées (formats, codes, langages différents, banques de données, sites, blogs, réseaux sociaux, terminaux connectés comme les smartphones, puces RFID, capteurs, caméras...). [3]
- **Volume** : le poids total des données collectées. [6]

Les 3 v du Big Data

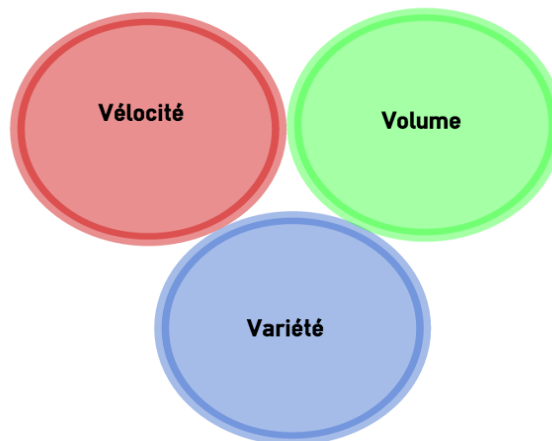


Figure1.1 Les 3 Caractéristiques des Big Data [22]

3.3. Les chiffres clés de la Big Data :

Nous créons chaque jour 2.500.000.000.000.000.000 (2.5 quintillion) octet de données. De quoi remplir 10 millions de disques blu-ray, qui empilée atteindraient la hauteur de 4 tours Eiffel. [14]

Nous pouvons bien constater que nous nageons dans un océan de donnée où le niveau de la mer augmente rapidement. [14]

3.4. Les enjeux du Big Data :

3.4.1. Les enjeux du Big Data en entreprise :

Pour les entreprises, le big data constitue une aide à la prise de décision au même titre que le business intelligence puisqu'il permet de mieux cerner les besoins des clients ou même d'anticiper leurs futures consommations. Les possibilités offertes sont vertigineuses et la plupart demeurent encore insoupçonnées pour l'heure. [2]

Toutefois, il existe de nombreuses applications concrètes, déjà accessibles ou qui devraient être proposées par les acteurs du marché à brève échéance :

- **Marketing** : La connaissance du client permise par le big data est tellement affinée qu'il devient envisageable pour l'entreprise de personnaliser la communication et de segmenter les offres promotionnelles avec une précision inégalée jusqu'ici. [2]
- **Production** : L'analyse des données peut être utilisée pour améliorer les processus, économiser les ressources énergétiques et naturelles, mieux gérer les stocks ou encore prévoir la maintenance des machines-outils. [2]
- **Commerce** : L'étude comportementale des clients basée sur le détail des commandes et la fréquence des achats peut servir à optimiser l'agencement des produits dans les rayons ou l'emplacement des points de vente et à fixer des prix de vente plus incitatifs. [2]
- **Recherche** : Le big data permet de traiter rapidement les masses d'informations issues des expérimentations, dans le but d'accélérer le prototypage et la mise sur le marché. [2]
- **Finance** : Les banques et les assureurs ont recours au big data pour déterminer le niveau de risque d'un client contractant un prêt ou une garantie et pour mettre en évidence un comportement suspect ou révéler des fraudes potentielles. [2]
- **Santé** : Le big data rend théoriquement possible une médecine préventive et personnalisée en lien avec des appareils connectés mesurant les données biométriques des patients et visant à leur proposer des conseils ou des traitements appropriés. [2]

- **Transport :** Le big data permet de modéliser les déplacements des usagers ou des salariés de l'entreprise afin d'optimiser les trajets et la fréquence des véhicules et donc d'accroître le taux de remplissage tout en faisant des économies de carburant. [2]

3.4.2. Enjeux des learning analytics :

Alors que le Big Data a investi les domaines de l'éducation et de la formation. L'enjeu des learning analytics est d'analyser les données d'apprentissage, mais surtout de les exploiter. Selon « Laurent Amice », Directeur du développement de la formation continue à l'Université de technologie de Troyes, « les learning analytics peuvent se définir comme les méthodes et outils d'analyse des données massives issues des apprentissages en ligne. » Concrètement, lors d'un apprentissage en ligne, les apprenants se connectent à une plateforme et suivent un parcours, ce qui génère des données : avancement, temps passé, chemin parcouru, pourcentage de bonnes réponses... Ce sont ces données qui une fois collectées et analysées, permettent de suivre et de comprendre le parcours des apprenants. [7]

3.4.3. Les learning analytics du côté du responsable de formation :

L'analyse des données d'apprentissage a pour finalité d'améliorer l'efficacité des dispositifs de formation, et donc l'impact de l'apprentissage. [7]

Pour les responsables de formation (enseignants, formateurs, responsables RH...), les bénéfices sont multiples :

- Suivre le parcours, les résultats et la progression des apprenants.
- Anticiper les difficultés pour intervenir auprès des apprenants.
- Mesurer l'impact de l'apprentissage et ajuster les cours et/ou les contenus.

Auxquels nous pouvons ajouter, pour l'entreprise :

- Identifier et anticiper les besoins de formation.
- S'assurer de la maîtrise des compétences requises.
- Mesurer le retour sur investissement de la formation.

3.4.4 Les learning analytics du côté de l'apprenant :

Pour sa part, l'apprenant peut suivre son parcours grâce aux données d'apprentissage. Mais pour lui, tout l'enjeu des learning analytics réside dans l'exploitation du Big Data pour bénéficier d'un apprentissage sur mesure. Car si l'analyse des données permet de comprendre

les parcours et de connaître les apprenants, il est ensuite possible de leur proposer des parcours d'apprentissage adaptés à leurs besoins, leurs préférences, leur rythme... [7]

Des algorithmes peuvent automatiser l'individualisation des parcours afin d'optimiser l'assimilation et la rétention des connaissances et/ou des compétences. L'adaptive learning (ou apprentissage adaptatif) permet donc de « conduire un apprenant vers le plus court chemin de la réussite ». [7]

3.5. Paysage technologique des Big Data :

Le paysage technologique autour des Big Data est complexe de par le nombre d'acteurs et la variété des technologies employées. Cependant, quel que soit la nature des données à traiter. [16]

Une solution applicative Big Data est généralement organisée selon le schéma ci-dessous.

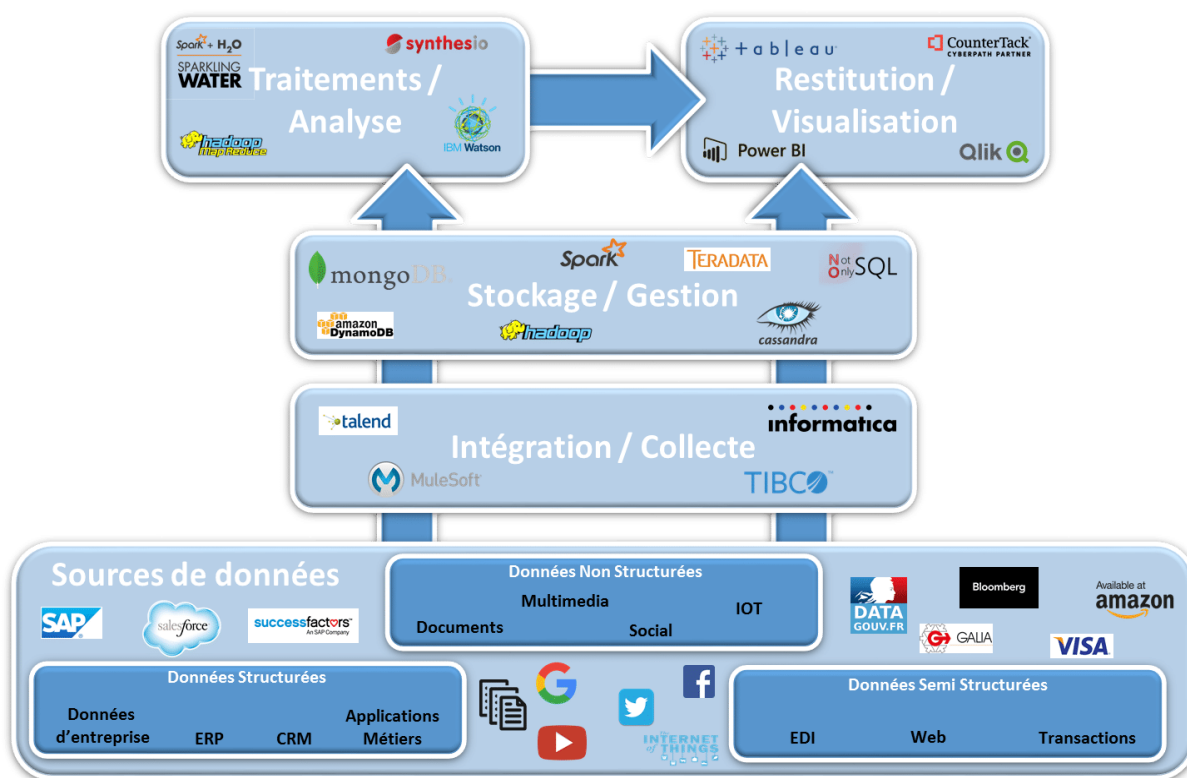


Figure1.2 Paysage technologique des Big Data [16]

(Les exemples sont donnés à titre indicatifs et ne sont pas exhaustifs des solutions)

3.5.1. Sources de données :

Elles sont diverses et variées. Elles peuvent être structurées et non structurées (données des capteurs, journaux Web, réseaux sociaux, documents ...). [16]

3.5.2. Intégration / Collecte :

Quel que soit la nature des données, elles devront être intégrées avant de pouvoir être traitées. Des traitements peuvent aussi être réalisés sur ces données lors de leur intégration afin d'en faciliter leur analyse ultérieure. [16]

3.5.3. Stockage / Gestion :

Afin de pouvoir être traitées, les données doivent pouvoir être stockées. Les approches « classiques » étaient principalement composées de bases de données structurées en SQL, comme Oracle par exemple. Dans une approche Big Data, le stockage et la gestion doivent pouvoir être réalisés sur un volume important de données avec une structure qui n'est plus celle de SQL. [16]

3.5.4. Traitement / Analyse :

La donnée une fois stockée doit être analysée afin d'en faire ressortir la « valeur ». C'est à cette étape qu'interviennent par exemple les technologies autour du machine learning, du search ou des analytics. [16]

3.5.5. Restitution/ Visualisation :

Une fois les informations extraites, il faut pouvoir les restituer. Soit au travers d'outils de visualisation ou bien d'outil plus métiers comme des outils marketing ou de contrôle de fraudes. [16]

3.6. Technologies de big data :

Cependant, les outils destinés à la collecte, le stockage et l'analyse doivent être adaptés pour tenir compte des nouvelles contraintes liées aux Big Data (les 3V). Les technologies associées commencent à se répandre dans leur usage ou à se développer. [16]

La liste ci-après ne se veut pas être exhaustive mais représente les quelques technologies les plus répandues.

3.6.1. MapReduce :

C'est un Framework de développement mis au point par Google en 2004 dont l'objectif est de permettre un traitement massivement parallèle sur les données. Il repose sur un principe de découpage des tâches à réaliser (grâce à la fonction Map), d'un traitement distribuer (grâce au Framework), puis d'un réassemblage des résultats (grâce à la fonction Reduce). Dans ce modèle, toute la partie relative à la distribution des traitements sur les différents serveurs est prise en charge par le Framework. Le développeur peut ainsi se concentrer uniquement sur le découpage et le réassemblage des données à traiter. [16]

3.6.2. Hadoop :

Comme MapReduce, Hadoop est un Framework pour le traitement de données massivement parallèle. Il possède 4 composants principaux. [16]

- Une bibliothèque et des utilitaires utilisés par tous les autres composants.
- Son propre système de gestion de fichiers distribués : le HDFS pour Hadoop Distributed File System.
- Son système de gestion des ressources : YARN.
- Enfin, pour la partie traitement, il est très proche de ce qui avait été produit par Google, C'est-à-dire les fonctions Map et Reduce.

La principale différence entre les deux Framework réside dans l'utilisation par Hadoop d'informations sur localisation des données (serveur, mémoire, fréquence, ...) permettant ainsi une optimisation des traitements. [16]

3.6.3. Spark :

Est un Framework qui se positionne comme le successeur de la combinaison MapReduce et Hadoop. Il est ouvert sur les différentes technologies à mettre en œuvre dans une architecture Big Data. Il peut par exemple reposer sur son propre moteur de gestion des clusters ou d'autres comme Yarn d'Hadoop. Et c'est la même chose concernant le système de gestion de fichiers distribués.

Spark intègre aussi nativement des fonctions permettant par exemple le traitement en continu de flux de données (Spark Streaming), le « machine learning » (MLlib) ou encore les calculs de graphes (GraphX). Dernier point, Spark peut travailler en mémoire vive, et donc réduire drastiquement les temps de traitement. Il est « jeune » mais dispose déjà d'une communauté énorme. [16]

3.6.4. NoSQL :

NoSQL a été pensé à l'origine dans le but de fournir un système de stockage et de manipulation de données non structurées en lignes et colonnes comme peut l'être une base de données « classique ».

Il existe différents types de base No SQL. Le choix est fait en fonction de la nature des données à manipuler (documents, son, images, ...). [16]

3.6.5. In Memory :

« In Memory » n'est pas une base de données à proprement parler mais des principes et mécanismes dont le but est de réaliser les traitements dans la mémoire de l'ordinateur et non plus uniquement sur des disques. Les temps d'accès à la donnée s'en trouvent donc

considérablement améliorés. De nombreuses bases de données sont maintenant « In Memory compatibles ». [16]

4. Conclusion :

Ce chapitre présente le Big Data, les différentes caractéristiques du Big Data, aussi on a focalisé sur l'historique du stockage des informations et les framework pour le traitement de donnée massive.