Projet Fin d'étude 1ere version

Amini Makhlouf Ben Aissa Tinhinane

2020/2021

Table des matières

Première partie

Etat de l'art

Chapitre 1

Big Data

Introduction

Avec la mise en place des services en ligne grâce à l'utilisation extensive d'Internet, le nombre de données générées qui transitent chaque jour sur le web, n'a fait que s'accroître de manière exponentielle, on parle ici de plus de 2,5 trillions d'octets générés quotidiennement, soit plus de 29.000 Giga-octets (Go) d'informations qui sont publiées dans le monde chaque seconde.

Ses données qui sont non pas que volumineuses mais aussi hétérogènes, viennent de toute part, la majeure partie de ses dernières proviennent de trois sources principales : les données sociales(les likes, les commentaires, les tweets, les photos/vidéos. . . etc), les données machines (Les capteurs tels que les appareils médicaux, les caméras routières, les satellites, les jeux et l'Internet des objets fournissent des données à haute vitesse, valeur, volume et variété) et les données transactionnelles (générées à partir de toutes les transactions quotidiennes qui ont lieu à la fois en ligne et hors ligne. Les factures, les ordres de paiement, les enregistrements de stockage, les reçus de livraison . . . etc).

Ses données qui sont utilisées par prêt de 6 milliards d'individus chaque jour, doivent être capturées, analysées, stockées, recherchées, partagées, visualisées, et transférées tout cela sans atteinte à la vie privée des utilisateurs, ce qui a poussé les chercheurs à trouver de nouvelles manières de réaliser tout cela étant donné que les outils traditionnels tels que le système de gestion de base de données relationnelles (SGBDR) et le SQL se retrouve dans l'incapacité de gérer ce nombre important et hétérogènes de données, et c'est ainsi qu'est né le "Big Data".

En effet, comme chaque domaine de connaissance, la terminologie naissante "Big Data" et la science des données sont utilisées pour parler de ce phénomène, Nous allons lors de ce chapitre présenter les concepts et les définitions se rapportant au domaine du "Big Data" quelques statistiques ainsi que, les intérêts, contraintes et caractéristiques de ce dernier.

1.1 Historique et quelques statistiques sur le Big Data :

L'expression «Big Data» serait apparue en octobre 1997 selon les archives de la bibliothèque numérique de l'Association for Computing Machinery (ACM), dans un article scientifique sur les défis technologiques à relever pour visualiser les «grands ensembles de données».

Il apparaît depuis fréquemment dans la presse et dans les revues universitaires, et des programmes de «Data Science» ont vu le jour dans le monde universitaire au cours des six dernières années. Le 29 mars 2012, WHOSTP a annoncé la "Big Data Research and Development Initiative" qui s'appuie sur des initiatives fédérales "allant de l'architecture informatique et des technologies de mise en réseau aux algorithmes, à la gestion des données, à l'intelligence artificielle, apprentissage automatique, développement et déploiement de cyber infrastructures avancées".

Au cours des six dernières années, au moins 17 programmes de science des données ont commencé dans les principales universités de recherche américaines et Internet regorge de publicités pour des livres et des cours de science des données.

Selon l'étude Data Age 2025, la sphère de données mondiale passera de 33 zettaoctets en 2018 à 175 Zo d'ici 2025. Près de 30% des données mondiales devront être traitées en temps réel et le stockage réalisé sur le Cloud public représentera 49% du volume total de données.

Pour ce qui est des statistiques le moins qu'on puisse c'est qu'elles sont impressionnants voici une figure qui permet de représenter la quantité de données générée en 60 secondes d'Internet en 2020 (Les données relevées ont été reprises de la compagnie Domo qui les a elle-même synthétisées à partir de nombreuses sources hétérogènes comme Business Insider, le New York Times, The Verge ou bien encore Hootsuite parmi d'autres et résumé par le site Visual Capitalist) :

FIGURE 1.1 – 1 minute d'Internet

En analysant la figure 1.1 on constate l'énorme quantité de données qui circule. En effet, durant cette minute sur internet alors que les utilisateurs de Facebook publient 147 000 photos, ceux d'Instagram partagent 347 222 stories et Twitter attire 319 nouveaux abonnés. Les plateformes de streaming ne sont pas en reste avec le SVOD Netflix notamment où les utilisateurs visionnent plus de 400 000 heures de vidéo en l'espace de 60 secondes et durant ce même laps de temps, 500 heures de vidéo sont publiées sur Youtube.

1.2 Définitions :

Mais alors qu'est-ce que le Big Data ?

Plusieurs définitions peuvent être données au Big Data, étant un objet complexe polymorphe, sa définition varie. Parmi elles nous citons :

Définition 1 : Le Big Data désigne l'ensemble des données numériques produites par l'utilisation des nouvelles technologies à des fins personnelles ou professionnelles. Cela regroupe les données d'entreprise ? Des contenus publiés sur le web, des transactions de commerce électronique, des échanges sur les réseaux sociaux, des données transmises par les objets connectés des données géolocalisées, ...etc.

Définition 2 : Le "Big Data" désigne les technologies et les initiatives qui impliquent des données trop diverses, en évolution rapide ou massives pour que les technologies, les compétences et les infrastructures conventionnelles puissent être traitées efficacement. Autrement dit, le volume, la vitesse ou la variété des données est trop important.

Définition 3 : Le terme Big Data fait référence aux données dont le coût de stockage, de gestion et d'analyse dans des systèmes de base de données traditionnels (relationnels et/ou monolithiques) serait généralement trop élevé. Habituellement, ces systèmes ne sont pas rentables, car ils ne disposent pas de la flexibilité nécessaire pour stocker des données non structurées (comme des images, du texte et des vidéos), pour accommoder des données "à haute vélocité" (en temps réel) ou pour s'adapter automatiquement à de très gros volumes de données (de l'ordre du pétaoctet).

FIGURE 1.2 – Le Big Data

1.3 Intérêts du Big-Data :

Dans tous les secteurs, les entreprises utilisent le Big Data engrangé dans leurs systèmes à différentes fins. Il peut s'agir d'améliorer les opérations, de proposer un meilleur service client, de créer des campagnes marketing personnalisées basées sur les préférences des consommateurs, ou tout simplement d'augmenter le chiffre d'affaires.

Grâce au Big Data, les entreprises peuvent profiter d'un avantage compétitif face à leurs concurrents n'exploitant pas les données. Elles peuvent prendre des décisions plus rapides et plus précises, s'appuyant directement sur les informations.

Par exemple, une entreprise peut analyser le Big Data pour découvrir de précieuses informations sur les besoins et les attentes de ses clients. Ces informations peuvent ensuite être exploitées pour créer de nouveaux produits ou des campagnes marketing ciblées afin d'accroître la fidélité client ou d'augmenter le taux de conversion. Une entreprise s'appuyant totalement sur les données pour aiguiller son évolution est qualifiée de " data-driven " (dirigée par les données).

On peut citer comme exemple : *Netflix, en effet En 2015, la lettre envoyée par Netflix à ses actionnaires a démontré que la stratégie Big Data portait ses fruits. Au premier trimestre 2015, 4,9 millions de nouveaux abonnés ont été enregistrés, contre quatre millions à la même période en 2014. De même, 10 milliards d'heures de contenu ont été diffusées pendant ce trimestre. Grâce à une utilisation intelligente du Big Data, l'influence de Netflix ne cesse de s'accroître.*

En outre, le Big Data est utilisé dans le domaine de la recherche médicale. Il permet notamment d'identifier des facteurs de risque de maladies, ou de réaliser des diagnostics plus fiables et plus précis. Les données médicales permettent aussi d'anticiper et de suivre les éventuelles épidémies.

Les mégadonnées sont utilisées dans presque tous les secteurs sans exception. L'industrie de l'énergie s'en sert pour découvrir des zones de forage potentielles et surveiller leurs opérations ou le réseau électrique. Les services financiers l'utilisent pour gérer les risques et analyser les données du marché en temps réel.

Les fabricants et les entreprises de transport, quant à eux, gèrent leurs chaînes logistiques et optimisent leurs itinéraires de livraison grâce aux données. De même, les gouvernements exploitent le Big Data pour la prévention du crime ou pour les initiatives de Smart City.

pour résumer, Le Big Data permet de construire de meilleurs modèles, qui produisent des résultats plus précis avec des approches extrêmement innovantes concernant la manière dont :

- Les entreprises se commercialisent et vendent leurs produits.
- La gestion des ressources humaines.
- La réaction au catastrophes naturelles.

Ces exemples ne sont finalement qu'une poignée des opportunités qu'offre le Big Data. Les entreprises, et pas seulement, devront faire preuve d'imagination, d'organisation et d'un énorme sens d'analyse pour prendre la pleine mesure du phénomène. De cette maîtrise découle de nouveaux usages qui bouleverse notre façon de concevoir Internet.

1.4 Les contraintes du Big Data :

L'intérêt du Big Data, c'est de pouvoir tirer profit de nouvelles données produites par tous les acteurs (les entreprises, les particuliers, les scientifiques et les institutions publiques) dans le but d'optimiser son offre commerciale, ses services, développer la recherche et le développement mais aussi créer des emplois. Il y a certes des avantages mais aussi des inconvénients du Big Data.

Certaines publications discutent des obstacles au développement d'applications de méga données. Les principaux défis sont énumérés comme suit :

- **Représentation des données :** De nombreux ensembles de données présentent certains niveaux d'hétérogénéité dans le type, la structure, la sémantique, l'organisation, la granularité et l'accessibilité. La représentation des données vise à rendre les données plus significatives pour l'analyse informatique et l'interprétation des utilisateurs. Néanmoins, une représentation incorrecte des données réduira la valeur des données originales et peut même empêcher une analyse efficace des données.
- **Réduction de la redondance et compression des données :** En général, il existe un niveau élevé de redondance dans les jeux de données. La réduction de la redondance et la compression des données sont efficaces pour réduire le coût indirect de l'ensemble du système en partant du principe que les valeurs potentielles des données ne sont pas affectées. Par exemple, la plupart des données générées par les réseaux de capteurs sont hautement redondantes.
- **Gestion du cycle de vie des données :** Par rapport aux progrès relativement lents des systèmes de stockage, la détection et le calcul omniprésents génèrent des données à des taux et des échelles sans précédent. Nous sommes confrontés à de nombreux défis urgents, dont l'un est que le système de stockage actuel ne peut pas supporter des données aussi massives. De manière générale, les valeurs cachées dans le Big Data dépendent de la fraîcheur des données.
- **Mécanisme analytique :** Le système analytique des méga données traitera des masses de données hétérogènes dans un temps limité. Cependant, les SGBDR traditionnels sont strictement conçus avec un manque d'évolutivité et d'extensibilité, ce qui ne pourrait pas répondre aux exigences de performance. Les bases de données non relationnelles ont montré leurs avantages uniques dans le traitement des données non structurées et ont commencé à se généraliser dans l'analyse des méga données. Même ainsi, il existe encore quelques problèmes de bases de données non relationnelles dans leurs performances et applications particulières. Des recherches supplémentaires sont nécessaires sur la base de données en mémoire et des échantillons de données basés sur une analyse approximative.
- **Confidentialité des données :** La plupart des fournisseurs ou propriétaires de services de méga données ne pouvaient actuellement pas maintenir et analyser efficacement des ensembles de données aussi énormes en raison de leur capacité limitée. Ils doivent s'appuyer sur des professionnels ou des outils pour analyser ces données, ce qui augmente les risques potentiels pour la sécurité. Par exemple, l'ensemble de données transactionnelles comprend généralement un ensemble de données d'exploitation complètes pour piloter les processus métier clés. Ces données contiennent des détails et certaines informations sensibles telles que les numéros de carte de crédit.
- **Gestion de l'énergie :** La consommation d'énergie des systèmes informatiques a beaucoup attiré l'attention du point de vue économique et environnemental. Avec l'augmentation du volume de données et des demandes analytiques, le traitement,

le stockage et la transmission de données massives consommeront inévitablement de plus en plus d'énergie électrique.

- **Expendabilité et évolutivité :** Le système analytique du Big Data doit prendre en charge les ensembles de données présents et futurs. L'algorithme analytique doit être capable de traiter des ensembles de données de plus en plus étendus et plus complexes.
- **Coopération :** L'analyse du Big Data est une recherche interdisciplinaire, qui nécessite la coopération d'experts dans différents domaines pour exploiter le potentiel du Big Data. Une architecture de réseau Big Data complète doit être mise en place pour aider les scientifiques et les ingénieurs dans divers domaines à accéder à différents types de données et à utiliser pleinement leur expertise, afin de coopérer pour atteindre les objectifs analytiques.

1.5 Caractéristiques des systèmes Big Data :

Les méga-données sont un terme générique utilisé pour désigner toute collection de données volumineuse et complexe qui peuvent dépasser la capacité de traitement des systèmes et techniques de gestion de données conventionnels. Les applications du Big Data sont infinies.

Les méga-données sont souvent caractérisées par le volume extrême des données, la grande variété de types de données et la vitesse à laquelle les données doivent être traitées. (Ces caractéristiques sont dites les 3V)

Ces caractéristiques ont été identifié pour la première fois par l'analyste Douglas Laney's membre du Gartner 10 dans un rapport publié en 2001. Plus récemment, plusieurs autres caractéristiques (autres V) ont été ajoutées aux descriptions des méga-données, notamment la véracité, la valeur et la variabilité. Bien que les méga-données ne correspondent à aucun volume de données spécifique, le terme est souvent utilisé pour décrire des téraoctets, des péta-octets et même des exa-octets de données capturées au fil du temps.

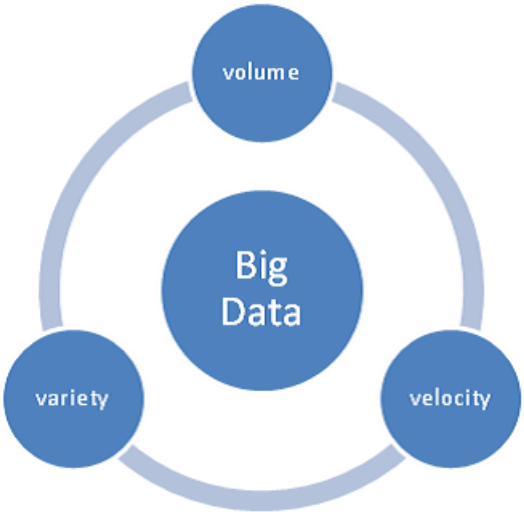

FIGURE 1.3 – Les 3 V.

Certaines personnes attribuent encore plus de V aux Big Data, les data scientiste et les consultants ont créé diverses listes contenant entre sept et 10 V. On donne dans ce qui suit 10 caractéristiques "10V" sur les méga-données sachant que Les 3 premier critères basiques du Big Data sont le volume la vitesse ainsi que la variété :

1.5.1 Le Volume :

Le caractère volume est certainement celui qui est le mieux décrit par le terme Big de l'expression Big Data. Volume fait référence à la quantité d'informations, trop volumineuse pour être acquise, stockée, traitée, analysée et diffusée par des outils standards. Ce caractère peut s'interpréter comme le traitement d'objets informationnels de grande taille ou de grandes collections d'objets.

Exemple : *les utilisateurs d'Instagram partagent 347 222 stories en 60secondes.*

1.5.2 La Variété :

Elle fait référence aux différentes formes toujours croissantes que les données peuvent prendre, en effet les données Big Data ne sont pas seulement des nombres, des dates et des chaînes. Les méga-données englobent également une grande variété de types de données, y compris des données structurées dans des bases de données SQL et des entrepôts de données, des données non structurées, telles que des fichiers texte et document conservés dans des clusters Hadoop ou des systèmes NoSQL, et des données semi-structurées, telles que des journaux de serveur Web ou la diffusion des données à partir de capteurs.

Exemple : *Un projet d'analyse des méga-données peut tenter d'évaluer le succès d'un produit et les ventes futures en corrélant les données de ventes passées, les données de retour et les données de révision des acheteurs en ligne pour ce produit.*

1.5.3 La vélocité :

Dernière dimension, tout aussi importante que les précédentes, la vélocité traduit la capacité à produire rapidement les données et à les transformer en temps utile pour leurs utilisateurs. L'exercice, déjà difficile dans un contexte "classique", prend toute sa valeur lorsqu'il doit être applique à d'immenses volumes de données de toutes sortes.

Exemple : *Google traite en moyenne plus de "40 000 requetés de recherche par seconde", ce qui représente environ 3,5 milliards de recherches par jour.*

1.5.4 La Véracité :

Elle fait référence aux biais, au bruit et aux anomalies dans les données. Ou, mieux encore, il fait référence aux incertitudes et à la fiabilité des données souvent incommensurables.

Exemple : *Dans le cadre d'un sondage réalisé par IBM, 27% des entreprises interrogées avouent ne pas être certaines de l'exactitude des données qu'elles collectent. De même, un chef d'entreprise sur trois utilise les données pour prendre des décisions, mais n'a pas vraiment confiance. Ce manque de véracité et de qualité des données coûte environ 3,1 trillions de dollars par an aux États-Unis.*

1.5.5 La Valeur :

Toutes les données collectées n'ont pas une valeur commerciale réelle et l'utilisation de données inexactes peut affaiblir les informations fournies par les applications d'analyse. Il est essentiel que les organisations utilisent des pratiques telles que le nettoyage des données et confirment que les données sont liées à des problèmes commerciaux pertinents avant de les utiliser dans un projet d'analyse de Big Data.

On peut dire que les autres caractéristiques du Big Data n'ont pas de sens si on ne tire pas de valeur commerciale de ces données. Les Données massives offrent une valeur substantielle : comprendre mieux les clients. Les cibler en conséquence, optimiser les processus

et améliorer les performances de la machine ou de l'entreprise. Avant de se lancer dans une stratégie Big Data, on doit comprendre le potentiel et les caractéristiques les plus difficiles.

Exemple : *La mise en place d'une analyse Big Data a permis à la société de développement d'éoliennes Vestas 11 d'optimiser son processus d'identification des meilleurs emplacements pour implanter ses éoliennes. Le traitement Big Data a engendré une augmentation de la performance de production d'électricité et une réduction des coûts énergétiques associés.*

Remarque : Certaines personnes attribuent encore plus de V aux Big Data ; les scientifiques des données et les consultants ont créé d'autres listes en ajoutant la variabilité, la validité, la visualisation, la volatilité ainsi que la vulnérabilité.

1.5.6 La Variabilité :

La variabilité dans le Big Data fait référence à plusieurs sens. Dans un premier temps elle désigne le nombre d'incohérences dans les données. Celles-ci doivent être détectées par des techniques de détection d'anomalies et de valeurs aberrantes pour faciliter la création d'analyse significative.

Les méga-données sont également variables en raison de la diversité de dimensions résultant de multiples types et sources de données. La variabilité peut également faire référence à la vitesse incohérente à laquelle les données volumineuses sont chargées dans la base de données.

Exemple : *L'équipe d'IBM 12 fait participer Watson 13 au célèbre jeu télévisé américain Jeopardy, un jeu ou les candidats doivent trouver les réponses à des questions posées. Watson devait "être capable de comprendre l'énoncé des questions, buzzer pour prendre la main, disséquer une réponse dans son sens pour déterminer quelle était la bonne question". Les mots n'ont pas de définitions statiques et leur signification peut varier énormément dans le contexte.*

1.5.7 La Validité :

Similaire à la véracité, la validité fait référence à la précision et à la correction des données pour l'usage auquel elles sont destinées. Selon Forbes 14, environ 60% du temps d'un scientifique est consacré au nettoyage de ses données avant de pouvoir effectuer une analyse. L'avantage de l'analyse des données massives est aussi primordial que celui des données sous-jacentes. On doit donc avoir de bonnes pratiques de gouvernance des données pour garantir une qualité des données cohérente, des définitions communes et des métadonnées.

Exemple : *La date d'une transaction est 02/07/1994 alors que l'activité de la société a débuté en 2000.*

1.5.8 La Volatilité :

On se pose les questions : 'quel âge doivent avoir les données pour qu'elles soient considérées comme non pertinentes, historiques ou obsolète ?', 'Combien de temps faut-il conserver les données ?' Avant l'ère du Big Data, en général, on stockait les données indéfiniment. Quelques téraoctets de données ne pouvaient pas engendrer de dépenses de stockage élevées.

En raison de la vitesse et du volume de ces données massives, leur volatilité doit être soigneusement prise en compte. Il est maintenant fondamental d'établir des règles pour la disponibilité et à la mise à jour des données a de garantir une récupération rapide des informations en cas de besoin.

Exemple : *Une entreprise e-commerce peut ne pas souhaiter conserver un historique des achats client d'un an. Parce qu'après un an la garantie par défaut sur leur produit expire, il n'y a donc aucune possibilité de restaurer ces données.*

1.5.9 La Visualisation :

Une autre caractéristique du Big Data est la difficulté à les visualiser. Les logiciels de visualisation de données volumineuses actuels sont confrontés à des problèmes techniques en raison des limitations de la technologie en mémoire, de leur faible évolutivité, de leur fonctionnalité et de leur temps de réponse. Il est impossible de se fier aux graphiques traditionnels lorsqu'on essaye de tracer un milliard de points de données. Il est donc nécessaire d'avoir différentes manières de représenter des données. Telles que la mise en cluster de données ou l'utilisation de cartes d'arbres, de sunbursts, de coordonnées parallèles, de diagrammes de réseau circulaires ou de cônes.

Si on associe cela avec la multitude de composante résultant de la variété et de la vélocité des données massives et des relations complexes qui les lient, il est possible de voir qu'il n'est pas si simple de créer une visualisation significative.

Prenons l'exemple : *du tableau suivant qui fait apparaître deux séries de chiffres : le chiffre d'affaires en France et le chiffre d'affaires du reste du monde. La lecture de ce tableau et sa signification ne sont pas immédiates.*

kEur	Janv	Fevr	Mars	Avril	Mai	Juin	Juillet	Août	Sept	Oct	Nov	Déc
France	10000	12000	14000	13000	15444	17028	15000	15804	18000	17500	19000	20958
Reste du monde	3444	3816	4038	3558	3864	4074	3558	2000	3594	3498	3612	4140
	13444	15816	18038	16558	19308	21102	18558	17804	21594	20998	22612	25098

FIGURE 1.4 – Évolution du chiffre d'affaires par région.

Mais si nous représentons les séries de chiffres sous forme graphique (ci-dessous), on comprend en un coup d'œil que le chiffre d'affaires en France progresse et que le chiffre d'affaires du Reste du monde stagne.

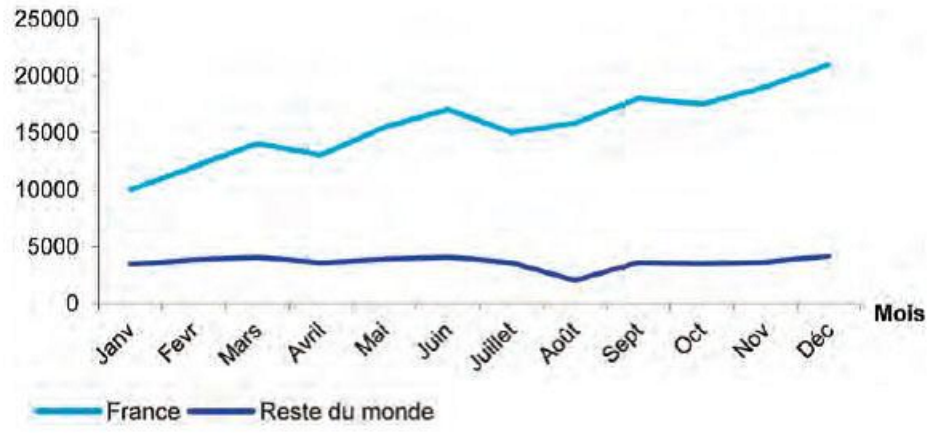

FIGURE 1.5 – Évolution du chiffre d'affaires par région.

1.5.10 La Vulnérabilité :

Le Big Data apporte de nouveaux problèmes de sécurité. Malheureusement, il y a quotidiennement des violations de données massives.

Exemple : *Rapporté par CRN15 : en mai 2016, un pirate informatique appelé Peace a posté des données sur le dark web pour les vendre, qui auraient inclus des informations sur 167 millions de comptes LinkedIn et 360 millions d'e-mails et de mots de passe pour les utilisateurs de MySpace 16.*

FIGURE 1.6 – Les 10 V.

Remarque : Il existe d'autres sources qui parlent de 54 Vs pour les caractéristiques du Big Data tel que la venue (le lieu), valence, vocabulaire, imprécision...etc.

1.6 Choses qui viennent du Big Data (Exemples du Big Data) :

Le concept de big data est une gestion groupée de différentes formes de données générées par divers appareils (Android, iOS, etc.), d'applications (applications musicales, applications Web, applications de jeux, etc.) ou d'actions (recherche par SE, navigation à travers des types similaires de pages Web, etc.). Voici la liste de certains champs de données couramment trouvés qui sont sous l'égide du Big Data :

- **Données sur les boîtes noires :** Les données des boîtes noires sont un type de données recueillies à partir d'hélicoptères, d'avions et de jets privés et gouvernementaux. Ces données comprennent la capture des sons de l'équipage de conduite, l'enregistrement séparé du microphone ainsi que des écouteurs, etc.
- **Données boursières :** Les données boursières comprennent diverses données préparées sur « l'achat » et la « vente » de différentes décisions brutes et bien prises.
- **Données sur les médias sociaux :** Ce type de données contient des informations sur les activités des médias sociaux qui incluent des messages soumis par des millions de personnes dans le monde entier.
- **Données sur les transports :** Les données sur les transports comprennent les modèles de véhicules, la capacité, la distance (d'une source à l'autre) et la disponibilité de différents véhicules.
- **Données des moteurs de recherche :** récupérez une grande variété d'informations non traitées stockées dans les bases de données SE.

1.7 Les sources du Big Data :

Une approche Big Data permet d'enrichir les données de l'entreprise avec celles de sources externes. Il n'est pas ici question de tendre vers « l'infobésité » (en accumulant toujours plus d'information non exploitée) mais plutôt de se donner de nouveaux angles de vue sur l'activité de l'entreprise, la conjoncture dans son secteur, ou encore son positionnement sur le Web. Le Big Data s'appuie sur quatre sources de données :

1.8 Les métiers du Big Data :

Le marché du big data est à l'origine d'un nombre croissant de métiers Les entreprises de tous les secteurs cherchent désormais à exploiter les données à leur disposition pour aiguiller leur stratégie et leur développement. Toutefois, pour être en mesure d'exploiter ces données, les entreprises doivent s'appuyer sur des compétences et du savoir-faire de professionnels hautement qualifiés capables d'utiliser les technologies analytiques. Ainsi, le Big Data a donné naissance à de nombreux nouveaux métiers :

1.8.1 LE CHIEF DATA OFFICER

Le chief data officer se charge de gouverner la data, qui constitue un capital vital pour l'entreprise. Il a pour mission de trier les masses de données disponibles afin de faciliter l'accès à l'information pertinente permettant des prises de décision adaptées. Pour ce faire, il doit constamment vérifier la fiabilité des informations recueillies et s'appuyer sur des éléments objectifs provenant de données statistiques. Le chief data officer intervient dans la construction et la mise en application d'une stratégie de gouvernance des données et travaille en collaboration avec d'autres professionnels tels que les data scientists, les spécialistes en business intelligence et les statisticiens.

1.8.2 LE DATA ENGINEER

Le data engineer (ou ingénieur de données) est un professionnel spécialisé dans la gestion des données. Sa mission principale consiste à recueillir, croiser, trier et réaliser des opérations de nettoyage des données. Il doit aussi gérer leur stockage dans différentes bases de données et exploiter des masses d'information sous divers formats.

1.8.3 LE DATA SCIENTIST

Le data scientist, appelé également analyste en big data, est un spécialiste de l'analyse des données massives. Sa mission prend effet après celle de l'ingénieur de données : le data engineer intervient dans la gestion des données alors que le data scientist assure le traitement de ces données pour en extraire de la valeur. Pour ce faire, le data scientist se charge de développer des algorithmes statistiques afin de tirer des informations pertinentes permettant de classifier des données, d'anticiper un comportement ou encore de préconiser des actions appropriées. Il doit donc avoir de solides connaissances en informatique, en statistiques et en management. Le data scientist doit également maîtriser les techniques du datamining et les outils de traitement des bases de données tels que Hadoop, MapReduce, Java, BigTable et NoSQL.

Les analystes en big data interviennent dans divers domaines d'activité : ils développent dans l'e-commerce et les réseaux sociaux les algorithmes de recommandation de pages, de profils et de produits

1.8.4 L'ARCHITECTE BIG DATA

Data architect en anglais, l'architecte big data a pour mission principale d'organiser des données brutes. C'est un métier plus conceptuel que technique, qui assure la création des infrastructures de stockage et conçoit des solutions de gestion des données massives. Il propose également aux décideurs la cartographie des outils Hadoop à mettre en place. L'architecte big data travaille en étroite collaboration avec le data scientist, tout en lui fournissant les données brutes à traiter. Il intervient également dans l'étude de la faisabilité technique et la mise en place des outils et la configuration des machines.

1.8.5 LE DÉVELOPPEUR BIG DATA

Le développeur big data maîtrise les différents langages informatiques notamment Java et Python. Il assure la cohérence du système, la gestion des pannes et garantit la continuité du service. Les données massives sont en effet au centre des préoccupations du métier de développeur big data. Ce profil travaille aussi en collaboration avec le data scientist : alors que ce dernier intervient dans la conception des algorithmes facilitant la prise de décision, le développeur big data assure leur mise en marche. Il fait partie des rares profils du big data à pouvoir gérer toutes les catégories des outils d'Hadoop pour des objectifs d'évaluation.

1.8.6 LE GROWTH HACKER

Le growth hacker n'est pas simplement un métier, c'est surtout un état d'esprit qui permet de développer plusieurs techniques webmarketing. C'est un profil à la croisée du marketing, du développement logiciel et du big data, qui a pour mission d'accélérer la croissance d'un produit ou d'un service propre à la structure qui l'embauche. Il utilise pour cela des solutions digitales innovantes et des pratiques de pointe afin d'accroître le revenu de son entreprise. Pour ce faire, le growth hacker cherche à développer, à partir d'Hadoop, de nouveaux produits et de nouvelles fonctionnalités. Il utilise également les outils de base de données (SQL) et les langages d'abstraction. De plus, comme tous les professionnels du marketing, il est en recherche constante de clients. Le growth hacker est très prisé par les start-up et les entreprises qui souhaitent se réinventer constamment.

1.8.7 LE DATA MINER

Le data miner assure la transmission des connaissances utiles à la progression de l'entreprise. Il dégage ainsi les tendances relatives à la consommation des clients pour en sortir avec une stratégie marketing réalisable sur le terrain. Pour se positionner sur le marché, il prend en compte les habitudes de consommation et les tarifs appliqués par la concurrence. De plus, il assure le tri des informations potentiellement exploitables, analyse les données après les avoir formatées et nettoyées. Il réalise aussi des rapports d'analyse, des tableaux de visualisation des données et compare les performances de l'entreprise pour les ajuster aux objectifs et prévisions. Le data miner a de grandes capacités d'observation et d'analyse de données.

1.8.8 L'ADMINISTRATEUR BIG DATA

L'administrateur joue un rôle primordial dans la structure informatique d'une entreprise. Il intervient dans la conception, l'optimisation et la configuration des infrastructures de stockage des données massives. Il assure également la sécurisation des données ainsi que l'attribution des autorisations et des droits d'accès aux différents utilisateurs. L'administrateur big data maîtrise les langages de programmation, les outils d'administration Hadoop et les protocoles de sécurité. Il vérifie la disponibilité de l'information à tout moment et apporte les modifications nécessaires sur les bases de données.

1.9 Les technologies du Big Data :

Cette technologie est importante pour présenter une analyse plus précise qui conduit l'analyste d'affaires à prendre des décisions très précises, assurant ainsi une efficacité opérationnelle plus considérable en réduisant les coûts et les risques commerciaux. Maintenant, pour implémenter de telles analyses et détenir une telle variété de données, il faut avoir besoin d'une infrastructure qui puisse faciliter et gérer et traiter d'énormes volumes de données en temps réel. De cette façon, le Big Data est classé en deux sous-catégories, le Big Data opérationnel qui comprend des données sur des systèmes et le Big Data analytique qui comprend des systèmes.

Nous décrivons ensuite ici tous les composants qui font partie des solutions Big Data sous de nombreux angles : matériel, méthodologies, logiciels et applications de base, etc.

Pour mieux catégoriser ces concepts, nous les avons a répartis en différentes sections selon l'objectif visé par chacun. Ces catégories sont : l'infrastructure, le stockage, le traitement et les composants de haut niveau.

1.9.1 Les infrastructures :

Le développement du Big Data commence avec les clusters Big Data 17 qui exécutent en parallèle les instructions d'un logiciel de haut niveau. Le cluster est artitionné en deux types de nœuds selon la fonction principale exercée :

- Nœuds de données ou esclaves (informatique).
- Nœuds de gestion ou maîtres (gestion).

Outre leur fonction, le maître et les esclaves peuvent être différenciés par leurs capacités de calcul et leur quantité dans le champ de nœuds.

Les esclaves sont chargés de surveiller les données partitionnées, de traiter et d'interroger les données locales. Les unités de données et de traitement doivent être aussi proches que possible pour éviter les retards introduits par les mouvements entre les partitions. Les nœuds de données sont gourmands en disque et standard en termes de capacités de calcul et de mémoire.

Les maîtres reçoivent et transforment les programmes des applications clientes en instructions parallèles qui peuvent être comprises par les esclaves. Une fois que les applications clientes ont atteint le démon maître, elles finissent par démarrer ou réveiller plusieurs processus dans les esclaves qui retournent finalement une sortie suivant la direction opposée. Parmi l'ensemble des responsabilités approuvées pour les nœuds de gestion figurent :

- La récupération après défaillance.
- La gestion des ressources.
- La planification des travaux.
- La surveillance ou la sécurité.

Pour accomplir ces tâches, les maîtres nécessitent une puissance de calcul et de mémoire élevée. Dans les clusters Big Data standard, il suffit de garder deux maîtres supports qui se surveillent mutuellement.

Les deux types de nœuds sont connectés via une connexion réseau, généralement LAN (Ethernet ou InfiniBand). Certaines configurations permettent également de connecter les maîtres de déférents centres de données sur un réseau WAN pour éviter facilement les défaillances du système. Dans chaque centre de données, le maître et les esclaves sont interconnectés en privé pour ingérer des données, déplacer des données entre les nœuds et effectuer des requêtes. Il existe également un autre réseau public qui sert de façade entre le client et le service de gestion (SSH, VNC, interface web,..)

1.9.2 Les technologies de traitements :

Dans cette section nous parlerons de l'arrivée des technologies de traitement ajustées, plus spécialement sur la mise au point de modes de calcul à haute performance (MapReduce), nous parlerons de (Hadoop) une solution de Big Data très largement utilisée pour effectuer des analyses sur de très grands nombres de données, et enfin nous clôturons cette section avec le développement de nouvelles bases de données adaptées aux données non structurées (NoSQL) et nous verrons qu'est-ce que le NewSQL].

1. **HADOOP :**

   Hadoop est un framework logiciel open source permettant de stocker des données, et de lancer des applications sur des grappes (cluster) de machines standards. Cette solution offre un espace de stockage massif pour tous les types de données, une immense puissance de traitement et la possibilité de prendre en charge une quantité de tâches virtuellement illimitée. Basé sur Java, ce framework fait partie du projet Apache, sponsorisé par Apache Software Foundation.

   Grâce au framework MapReduce, il permet de traiter les immenses quantités de données. Plutôt que de devoir déplacer les données vers un réseau pour procéder au traitement, MapReduce permet de déplacer directement le logiciel de traitement vers les données.

   Dans son principe, Hadoop se compose essentiellement de :

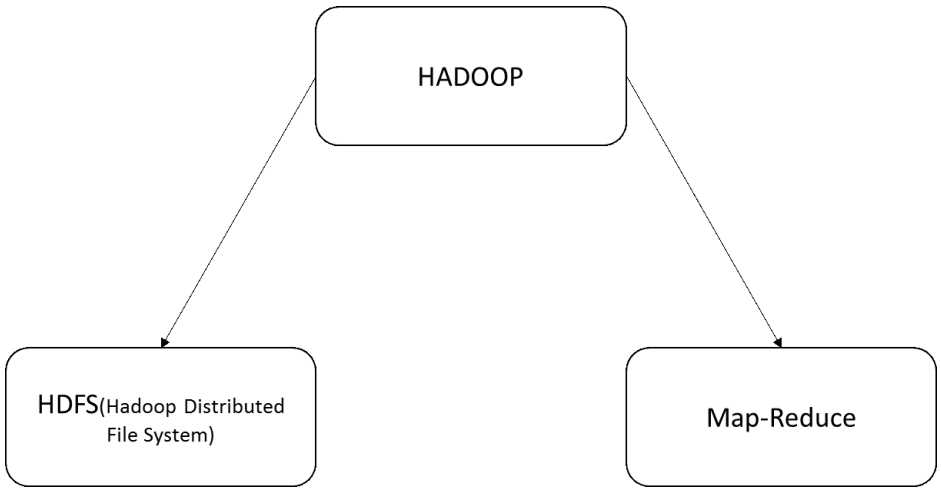

FIGURE 1.7 – Les composants d'Hadoop.

— **Système de gestion de fichiers HDFS (Hadoop Distributed File System) :** HDFS est un système de fichiers distribué, extensible et portable Inspiré par GFS et écrit en Java. Il est conçu pour être un système de stockage distribué, évolutif et résilient, conçu pour interagir facilement avec MapReduce. Il fournit une bande passante d'agrégation importante tout au long du réseau. Comme pour GFS, un réseau HDFS est composé d'un nœud maître appelé Namenode et des serveurs de données appelés Datanodes, de grande taille par défaut 64 Mo pour optimiser les temps de transfert et d'accès. Il est toutefois possible de monter à 128 Mo, 256 Mo, 512 Mo voire 1 Go.

— Ces blocs sont ensuite répartis sur plusieurs machines, permettant ainsi de traiter un même fichier en parallèle. Pour garantir une tolérance aux pannes, les blocs

de chaque fichier sont répliqués sur plusieurs machines. Notez que si la taille du fichier est inférieure à la taille d'un bloc, le fichier n'occupera pas la taille totale de ce bloc.

— **Modèle de programmation Map-reduce :** Le framework MapReduce permet de traiter les immenses quantités de données. Plutôt que de devoir déplacer les données vers un réseau pour procéder au traitement, MapReduce permet de déplacer directement le logiciel de traitement vers les données. Ce modèle fera l'objet de la deuxième technologie que nous allons présenter.

— Une collection d'outils spécifiques pour HDFS et Map Reduce comme des API et des frameworks.

2. **Map-reduce :**

MapReduce a été introduit par Google et décrit en détails dans la publication « MapReduce : Simplified Data Processing on Large Clusters » publiée en 2004 et ça pour faciliter la mise en œuvre de ses workfows de traitement parallèle. L'objectif principal était de remplacer la programmation complexe et non intuitive sur l'informatique distribuée (préalablement abordée par les plateformes HPC) par une plateforme transparente moderne avec seulement deux fonctions : Map et Reduce. Ces deux fonctions définies par l'utilisateur permettent aux utilisateurs d'utiliser les ressources distribuées sans se plaindre du réseau, de la planification, de la récupération après défaillance, etc.

Un modèle aussi intuitif que le MapReduce ne nécessite pas d'expertise concernant le parallélisme et les systèmes distribués. Son Framework Plug-and-Play embarque tous les détails pour implémenter les systèmes de calcul parallèle, la persistance et la résilience, l'optimisation et l'équilibre des ressources.

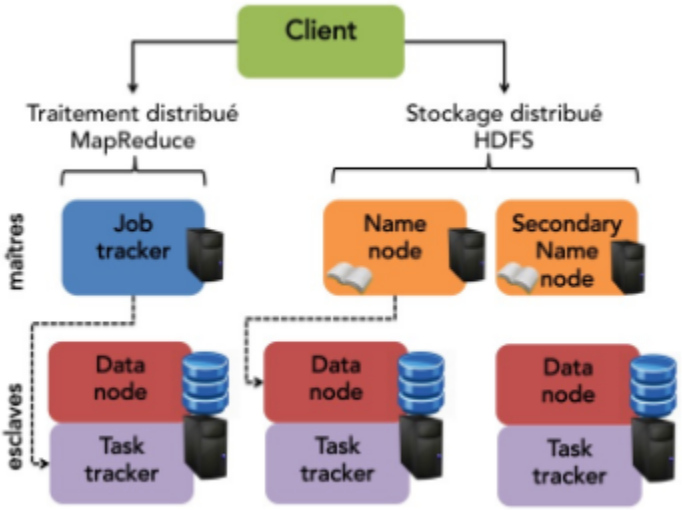

FIGURE 1.8 – L'environnement Hadoop.

— **Principe de MapReduce :**
Le Framework se décompose en deux parties. La fonction Map permet aux différents points du cluster distribué de distribuer leur travail. La fonction Reduce permet de réduire la forme finale des résultats des clusters en un seul résultat. Cela est rendu possible grâce au système de fichiers distribués HDFS d'Hadoop.

Ce Framework est également constitué de plusieurs composants

— **Job tracker :** Est le nœud principal qui gère toutes les tâches et les ressources d'un cluster.
— **Les TaskTrackers :** Sont les agents déployés sur chaque machine d'un cluster pour lancer la map et réduire les tâches.
— **JobHistoryServer :** est un composant permettant de suivre les tâches complétées, généralement déployé comme une fonction séparée ou avec JobTracker.

Exemple d'utilisation de MapReduce : *Dans ce qui suit, on vient mettre en clair le fonctionnement de MapReduce avec l'exemple typique Word Count schématisé afin de mieux cerner le rôle de chaque fonction de ce framework. S'il est possible de compter manuellement le nombre de fois qu'un mot apparaît dans un roman, cela prend beaucoup de temps. Si l'on répartit cette tâche entre une vingtaine de personnes, les choses peuvent aller beaucoup plus vite. Chaque personne prend une page du roman et écrit le nombre de fois que le mot apparaît sur la page. Il s'agit de la partie Map de MapReduce. Si une personne s'en va, une autre prend sa place.*
Cet exemple illustre la tolérance aux erreurs de MapReduce. Lorsque toutes les pages sont traitées, les utilisateurs répartissent tous les mots dans 26 boîtes en fonction de la première lettre de chaque mot. Chaque utilisateur prend une boîte, et classe les mots par ordre alphabétique. Le nombre de pages avec le même mot est un exemple de la partie Reduce de MapReduce.

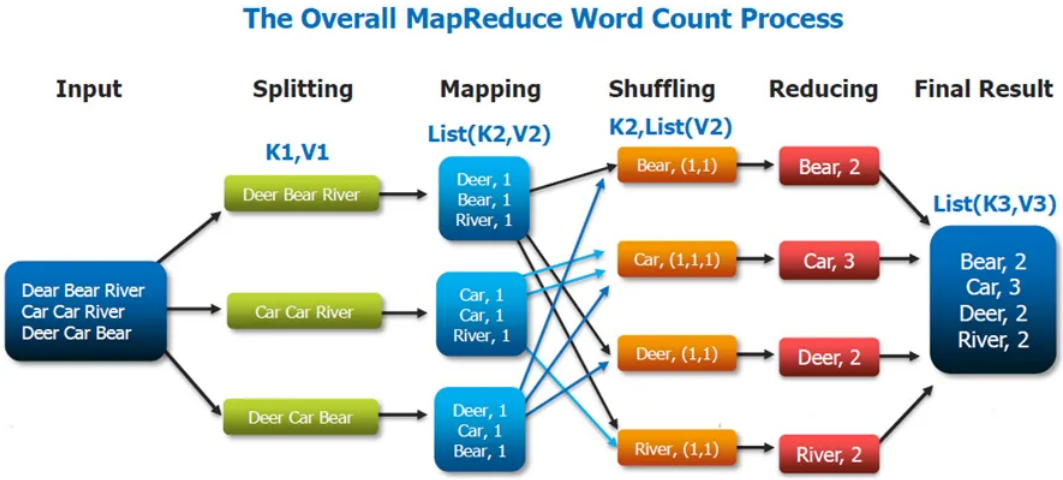

FIGURE 1.9 – Exemple de Word Count.

— **Les avantages de MapReduce :** Parmi les avantages de la programmation MapReduce nous citons
  — La scalabilité.
  — La flexibilité.
  — La sécurité et l'authentification.
  — Le traitement parallèle.
  — La disponibilité.
  — Un modèle simple de programmation.

3. **Les bases de données NoSQL :**

   De nos jours, l'ubiquité de la connexion Internet est une réalité (les voitures que nous conduisons, les montres que nous portons, nos petits appareils médicaux domestiques, nos réfrigérateurs et congélateurs, nos Smartphones et ordinateurs portables). De plus, les données numériques produites par les êtres humains, dont les séquences vidéo, les photos et autres, atteignent des volumes importants de plusieurs EO par jour. Ces données actuellement stockées dans des bases qui leur ont été conçues spécifiquement sont gérés par des logiciels de gestion de bases de données volumineuses, jouant le rôle d'intermédiaires entre les bases de données d'un côté et les applicatifs et leurs utilisateurs de l'autre. On parle ici des bases de données non-relationnelles, dites NoSQL.

   Concrètement une base de données NoSQL est une approche de la conception des bases et de leur administration particulièrement utile pour de très grands ensembles de données distribuées. Elle englobe une gamme étendue de technologies et d'architectures, afin de résoudre les problèmes de performances en matière d'évolutivité et de Big Data que les bases de données relationnelles ne sont pas conçues pour affronter. De plus elle est particulièrement utile lorsqu'une entreprise doit accéder, à des fins d'analyse, à de grandes quantités de données non structurées ou de données stockées à distance sur plusieurs serveurs virtuels du Cloud.

1.9.3 Les technologies de stockage :

1. **Stockage "In-Memory" :** Pour des analyses encore plus rapides, les traitements directement en mémoire sont une solution. Une technologie bien qu'encore trop coûteuse il est vrai pour être généralisée. Les bases de données "In Memory" sont généralement construites comme des base relationnelles. Elles sont conformes aux exigences ACID (Atomicity, Consistency, Isolation, Durability) qui garantissent l'intégrité des transactions. Les données contenues en mémoire sont volatiles par principe. Un système de sauvegarde périodique par image disque, snapshot, permet de sauvegarder la base. Ce système est complété d'une historisation des transactions afin de remettre la base en état en cas de coupure de courant.
2. **Le Cloud computing :** C'est une solution d'externalisation capable de louer de puissants moyens de calcul. Ceux-ci sont dotés de larges capacités de stockage extensibles et adaptés aux traitements des Big Data. Au cœur des réflexions sur les infrastructures IT, de nouvelles offres Big data as a Service (BDaaS).

   C'est un terme général employé pour désigner la livraison de ressources et de services à la demande par Internet. Il désigne le stockage et l'accès aux données par l'intermédiaire d'Internet plutôt que via le disque dur d'un ordinateur. Il s'oppose ainsi à la notion de stockage local, consistant à entreposer des données ou à lancer des programmes depuis le disque dur.

1.10 Challenge et future du Big Data :

1.10.1 Les challenge

La Croissance rapide des données crée un problème pour rechercher des idées qui l'utilisent. Il n'existe aucun moyen 100

- **Stockage :** La génération d'une quantité aussi massive de données a besoin d'espace pour le stockage, et les organisations sont confrontées à des défis pour traiter des données aussi étendues sans outils et technologies appropriés.
- **Données peu fiables :** Il n'est pas garanti que les données big data collectées et analysées sont totalement (100%) précis. Les données redondantes, les données contradictoires ou les données incomplètes sont des défis qui demeurent à l'intérieur.
- **Sécurité des données :** Les entreprises et les organisations qui stockent des données aussi massives (des utilisateurs) peuvent être la cible de cybercriminels, et il existe un risque de vol de données. Par conséquent, le chiffrement de ces données colossales est également un défi pour les entreprises et les organisations.

1.10.2 Le future du Big Data

Conclusion :

Chapitre 2

Le cloud computing :