



Université Panthéon Sorbonne

Mémoire Master 1 MIAGE

Pollution et changement climatique mondial, en quoi le Big data est un défi important ?



Réalisé par : Camille MEURIC

Encadré par : Nicolas HERBAUT

2020-2021

Table des matières

Préface	3
Remerciements.....	4
Introduction.....	5
Motivation et contexte	7
1_ Les origines du big Data	7
2_ L'impact de la technologie sur l'environnement	14
Etat de l'art.....	21
I_ La collecte des données.....	21
A) Vision générale	21
B) Discussion	24
II_ L'analyse des données	29
A) Vision générale	29
B) Discussion	35
III_ L'exploitation des données.....	39
A) Vision générale	39
B) Discussion	42
Conclusion	45
Références	47
Liste des tableaux et figures.....	49

Préface

Dans le cadre de mon master 1 MIAGE, je suis amenée à effectuer un mémoire de type état de l'art. L'objectif de ce mémoire est de définir une problématique et de faire une présentation pédagogique et synthétique des solutions existantes reposant sur des articles et revues scientifiques.

S'inscrivant dans le contexte actuel, je voulais allier l'informatique aux problématiques liées au changement climatique et à la pollution.

C'est à partir de cette idée que j'ai choisi d'écrire ce mémoire qui met en évidence les relations entre le Big Data et le changement climatique.

Remerciements

Je tiens tout d'abord à remercier sincèrement mon tuteur enseignant, monsieur HERBAUT Nicolas, qui m'a encadrée, guidée, conseillée et soutenue durant l'élaboration de mon mémoire.

Je remercie également les membres de l'équipe pédagogique de l'Université Paris 1 Panthéon-Sorbonne dont les enseignements m'ont permis d'acquérir, tout au long de ma formation, des connaissances aussi bien techniques que pédagogiques, qui m'ont été essentielles afin de mener à bien mon travail.

Introduction

Le changement climatique est une réalité. Le climat de la Terre change à une vitesse jamais vue dans l'histoire de l'humanité. La température moyenne de la Terre a augmenté de 1 °C au cours des 100 dernières années. On note que les années 2015 à 2020 ont été les plus chaudes jamais enregistrées.¹

Les changements climatiques entraînent aussi des changements sur les conditions météorologiques mondiales, les courants océaniques et d'autres systèmes. Nous sommes au cœur de cette actualité, ce n'est plus dans l'avenir que ces changements vont se dérouler, c'est maintenant. Les catastrophes s'enchaînent depuis plusieurs mois avec un taux d'inondations record, des vagues de chaleur et d'incendies. Même si chaque région ressent ces changements différemment, ils sont néanmoins bien présents. Le temps presse et il devient urgent de trouver des solutions.

C'est pourquoi le projet « Baltimore Open Air » est né. La ville de Baltimore se classe régulièrement comme ayant la pire qualité de l'air sur la côte Est des États-Unis, selon l'Agence fédérale de protection de l'environnement, où des questions sur les risques importants pour la santé de la communauté se posent.

Le but du projet est de construire et d'installer des capteurs sans fil en forme de cube autour de Baltimore afin de recueillir des données et de mesurer l'effet des espaces verts sur la chaleur à Baltimore.²

La question à laquelle le projet souhaite répondre est la suivante : pouvons-nous créer un réseau de capteurs à faible coût pour recevoir des données ?

L'idée est que l'équipe de chercheurs, avec l'aide des membres de la communauté, construisent les capteurs et, en partenariat avec le Bureau du développement durable de la ville, les placent dans divers quartiers de la ville. Les capteurs collecteront des données critiques sur les événements de canicule, la température et la pollution, et ces données seront

¹ Earth Science Communications Team.2021 [Global Climate Change](#). NASA

transmises en temps réel aux services cloud, prêtes à être analysées et utilisées par les chercheurs, les groupes communautaires et d'autres parties intéressées. Les données seront également utilisées pour mesurer l'efficacité des efforts écologiques tels que planter un plus grand nombre d'arbres, installer des toits blancs et d'autres infrastructures vertes.²

Il est grand temps de fournir aux citoyens et aux chercheurs les informations nécessaires pour prendre des décisions environnementales et sanitaires. Comme pour Baltimore, le Big Data pourrait être la solution au problème du réchauffement climatique et autres problèmes de climat qui sévissent sur la planète.

Cette nouvelle façon de penser les données environnementales permettrait de faire de grandes avancées, cependant l'utilisation de ces technologies ne sont pas sans dégâts sur l'environnement.

Ce qui nous amène à cette problématique : **Quel est l'impact du big data sur le changement climatique et la pollution ?**

Dans un premier temps, nous allons nous intéresser aux origines du Big Data dans sa globalité, avant d'aborder l'influence des nouvelles technologies sur l'environnement. Nous évoquerons dans un second temps, les enjeux de la collecte de données sur l'environnement, ainsi que les différentes techniques d'analyse de ces données avant de conclure sur l'exploitation de celles-ci.

² Kate Pipkin. 2017. Can Big data help solve big climate problems? Hub

Motivation et contexte

1_ Les origines du big Data

Le big data est défini de façons différentes selon les sources, c'est pourquoi nous allons remonter dans le temps et montrer l'évolution des définitions du big Data au fil de l'histoire.

C'est en 1944, que Fremont Rider, bibliothécaire de l'Université Wesleyan, en se basant sur son observation du fait que les bibliothèques des universités américaines doublent tous les 16 ans, a spéculé que la bibliothèque de Yale en 2040 aurait *"approximately 200,000,000 volumes, which will occupy over 6,000 miles of shelves... [requiring] a cataloging staff of over six thousand persons."*⁴

De 1944 à 1980, de nombreux articles et présentations ont observé *"l'explosion de l'information"* et le besoin de capacité de stockage.

C'est alors qu'en 1997, Michael Cox et David Ellsworth produisent le premier article de la bibliothèque numérique de l'ACM (Association for Computing Machinery) qui utilise le terme « Big Data ».

"Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of big data. When data sets do not fit in main memory (in core), or when they do not fit even on local disk, the most common solution is to acquire more resources." - Application-controlled demand paging for out-of-core visualization.⁴

En 2000, Francis Diebold établit un lien explicite entre le Big Data et la façon dont nous le comprenons aujourd'hui, dans l'article "Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting".

"Recently, much good science, whether physical, biological, or social, has been forced to confront—and has often benefited from—the "Big Data" phenomenon. Big Data refers to the

explosion in the quantity (and sometimes, quality) of available and potentially relevant data, largely the result of recent and unprecedented advancements in data recording and storage technology.”⁴

C'est en 2001, que Doug Laney, Meta Group (Gartner), a publié "3D Data Management: Controlling Data Volume, Velocity, and Variety". C'est à ce moment que la définition des 3 V pour Volume, Variété et Vitesse, par lequel le Big data est le plus souvent caractérisé est née.

Volume

Lorsque nous évoquons le big data, nous pensons tout de suite à l'important volume de données.

Cependant, il peut être compliqué de se représenter ce que signifie, méga, giga ou teraoctet. Par exemple, si nous prenions comme unité la seconde, alors une megaseconde représenterait 12 jours, une giga seconde, 32 ans et une teraseconde à 33 000 ans.

Ce flux de données créées, ne cesse d'augmenter chaque année, certains spécialistes estiment même qu'il sera multiplié par 45 d'ici à 2035. Il passerait ainsi de 47 à 2142 zettaoctets.³

³ Marjolaine Tasset. Le Volume de données mondial sera multiplié par 45 entre 2020 et 2035. JDN

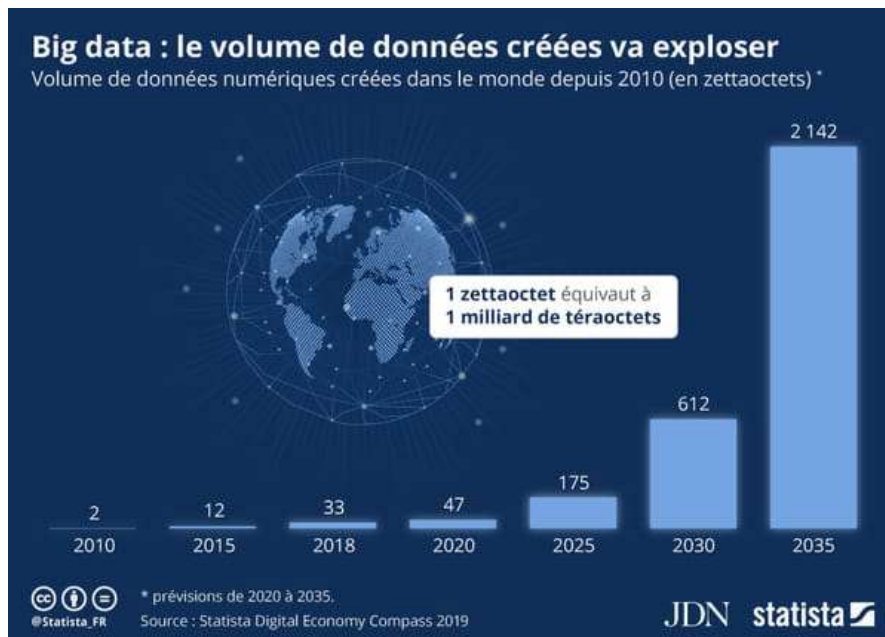


Figure 1: Statistique de l'évolution du volume de données créées³

Variété

La variété d'une donnée est caractérisée par les différents types de données disponibles.

L'augmentation des types de données provenant de sources telles que les médias sociaux, l'interaction machine à machine et les appareils mobiles a créé une énorme diversité au-delà des données de transaction traditionnelles. Les données ne s'intègrent plus dans une structure nette et facile à utiliser.

Vélocité

La vélocité décrit la fréquence à laquelle les données sont à la fois générées, capturées, mises à jour et analysées. En raison des récents développements technologiques, non seulement les consommateurs, mais aussi les entreprises génèrent plus de données en moins de temps.

À cette vitesse, les entreprises ne peuvent utiliser ces données que si elles sont collectées et partagées en temps réel. C'est à ce stade que de nombreux systèmes d'analyses, de CRM (*customer relationship management*), de personnalisation, de point de vente et d'autres échouent.

Pour finir, 2005 est l'année au cours de laquelle Tim O'Reilly a publié son article révolutionnaire intitulé *"What is Web 2.0 ? "*. Il a explicitement utilisé le terme "Big Data" pour désigner un vaste ensemble de données qu'il est presque impossible de gérer et de traiter à l'aide des outils de veille stratégique traditionnels. Il s'agit sans aucun doute de la forme actuelle de définition du Big Data la plus répandue.⁴

Même s'il n'existe pas de définition unique du Big Data qui soit universellement acceptée, il y a quelques concepts communs sur lesquels presque tous semblent s'accorder.

- Ce sont des données de grand volume (> 1 PetaOctets).
- Elles ne sont pas d'un seul type, c'est-à-dire structurées et une variété de données structurées, non structurées, etc.
- Elles sont générées à un rythme beaucoup plus rapide que les données du passé à partir de toutes sortes de sources, y compris les médias sociaux.
- De nouvelles méthodes de stockage, de traitement, d'analyse, de visualisation et d'intégration sont nécessaires.

Comme nous l'énonçons rapidement dans les concepts, le stockage des données doit faire face à un renouvellement constant. Les bases de données relationnelles classiques ne peuvent pas gérer de gros volumes de données. Les nouveaux modèles de représentation, eux, garantissent les performances des volumes concernés.

Ces technologies, appelées Business Analysis and Optimization (BAO), permettent de gérer des bases de données massivement parallèles. Les acteurs de ce marché fournissent des modèles architecturaux («Big Data Architecture Framework», BDAF), comme MapReduce créé par Google et utilisé dans le framework Hadoop.

⁴ Gil Press. 2013. A Very Short Story Of Big Data. Forbes.

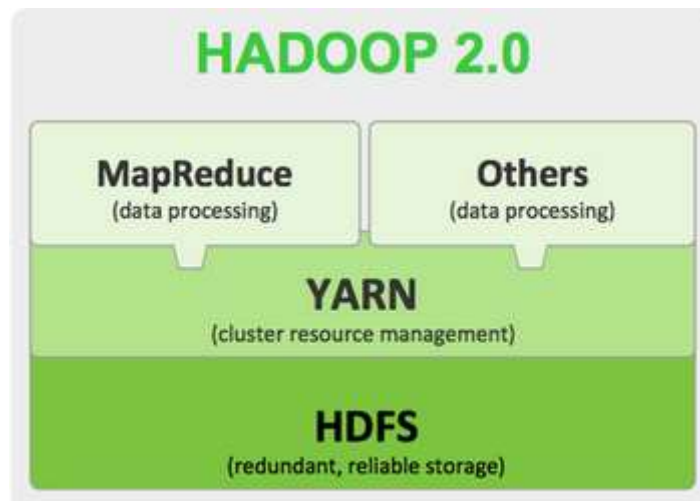


Figure 2: Hadoop's System⁵

Avec ce système, les requêtes sont séparées et distribuées à des nœuds parallélisés, puis exécutées en parallèle. Les résultats sont ensuite rassemblés et récupérés. Les acteurs du marché s'appuient sur des systèmes à haut niveau d'évolutivité et des solutions basées sur NoSQL comme MongoDB ou Cassandra, au lieu des bases de données relationnelles traditionnelles.

⁵ Bastien L. 2018. Hadoop – Tout savoir sur la principale plateforme Big Data. Le Big Data

Etape 1	Etape 2	Etape 3	Etape 4	Etape 5	Etape 6
Document d'entrée à traiter	Le document est stocké et partitionné en plusieurs nœuds	Chaque nœud va exécuter la fonction MAP pour avoir un stockage formaté sous (clé, valeur)	Les mots sont regroupés selon la même clé	Chaque nœud va exécuter sur les données la fonction REDUCE pour combiner les valeurs des mêmes clés	Document de sortie

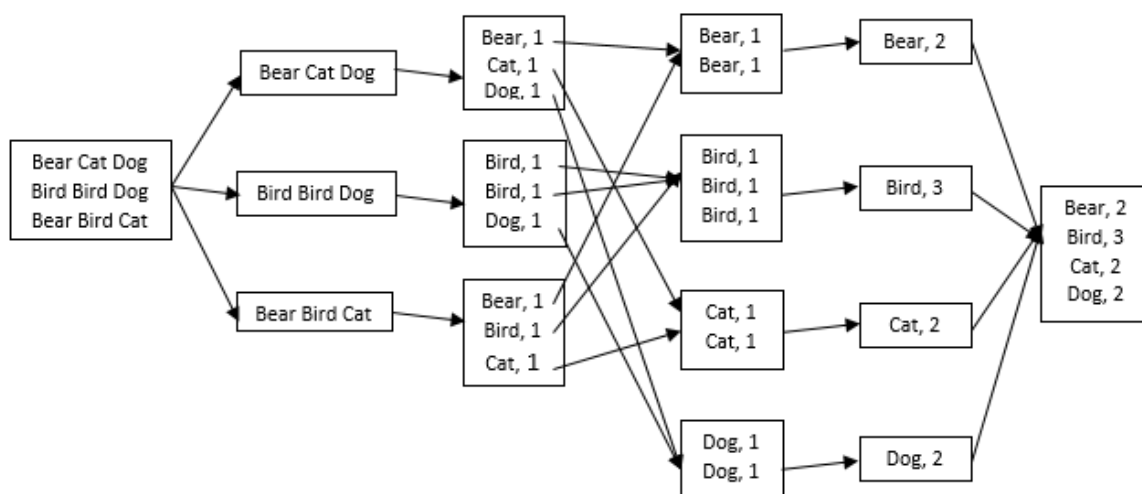


Figure 3 : Exemple d'utilisation du modèle Map Reduce, afin de compter le nombre d'occurrence d'un mot dans un texte donné⁶

Pour répondre aux problématiques du big data, l'architecture de stockage des systèmes doit être repensée et les modèles de stockage doivent se multiplier en conséquence.

Plusieurs systèmes ont émergé comme :

⁶ Onesmus Mbaabu.2020. Understanding Map Reduce in Hadoop. Section

- **Data Lake** : C'est une méthode de stockage de grandes quantités de données dans un cluster, qui permet la conservation de son format d'origine afin que tout format de données puisse être stocké rapidement et à moindre coût.

- **Cloud Computing** : Permet un accès via le réseau, accès aux services à la demande et accès aux services en libre-service via des ressources informatiques partagées et configurables.

Après avoir défini les problématiques de stockage du Big data, nous allons évoquer rapidement quelles utilisations nous pouvons faire de toutes ces données en les analysant et en quoi elles sont importantes dans différents domaines.

L'analyse du Big data permet par exemple :

- Comprendre les besoins des individus et les contraintes des usagers ;
- Adapter les infrastructures, réseaux et services publics en fonction de leur utilisation ;
- Assister la prise de décision des différents acteurs économiques ;
- Analyser et d'anticiper les comportements des consommateurs (analyse prédictive) ;
- Améliorer les performances, prévenir les pannes et maintenir les machines et appareils;

Le BigData a un large panel d'applications dans différents domaines : d'une part, il peut conduire des analyses croisées et prédictives dans la connaissance et l'évaluation, il peut effectuer des analyses de tendances dans le domaine du climat, de la politique environnementale ou sociale. D'autre part, il permet aussi d'évaluer la gestion des risques pour des entreprises, des assurances et pour l'industrie. Le Big data joue un rôle aussi dans les aspects suivants : la prise de décision politique, la sécurité et la lutte contre les activités criminelles, mais aussi dans la médecine lors de la compréhension du fonctionnement du cerveau.

Concernant son application au profit de l'écologie, le big data contient des données mondiales déterminantes « pour résoudre l'équation climatique ». Notamment pour améliorer l'efficacité énergétique des villes et bâtiments, pour les réseaux intelligents, pour vérifier l'application de réglementations visant à lutter contre la déforestation, la surpêche, la

dégradation des sols, ou à mieux gérer les déchets, éco-consommer. Cependant, il existe des points néfastes à cette évolution.⁴

2_ L'impact de la technologie sur l'environnement

Aujourd'hui, les nouvelles technologies font partie de notre quotidien. Se passer de son téléphone portable, ordinateur ou même Internet, est devenu difficile voir inconcevable. Selon les estimations des Nations Unies, d'ici 2035, la planète entière devrait pouvoir accéder normalement à Internet.⁷

Cependant, le lien établi entre ces technologies et l'environnement n'est pas très positif. Les dommages à l'environnement causés par la nouvelle approche sont en augmentation. Selon une étude publiée par GreenIT, la fabrication et l'utilisation de nouvelles technologies sont la principale source de consommation d'énergie et donc la principale source d'émissions de gaz à effet de serre.

Effet de serre, comment ça marche ?

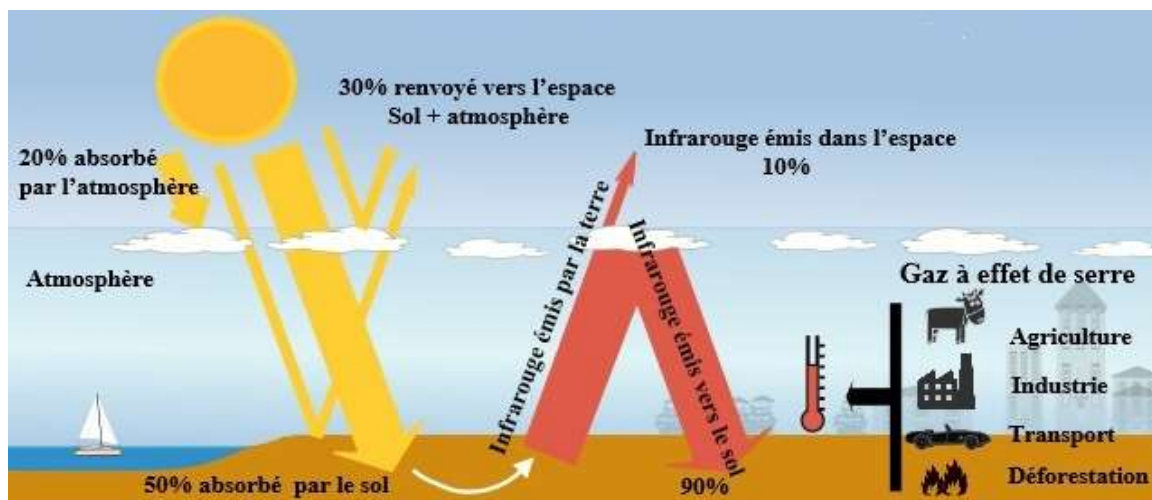


Figure 4: Représentation de l'effet de serre ⁸

⁷ Inconnu.2020. L'impact de la technologie sur l'environnement. Ellcie Healthy

Lorsque la Terre est éclairée par le Soleil, sa surface ré-émet vers l'espace une partie du rayonnement. Sauf que les gaz à effet de serre comme le CO₂ le méthane ou le protoxyde d'azote retiennent une partie de ce rayonnement infrarouge émis par la Terre et lui renvoient, ce qui contribue à la réchauffer.⁸

Or, c'est cette mécanique que l'homme a dérégulée. Les causes sont diverses : combustion de pétrole, déforestation et agriculture intensive, qui émettent de grandes quantités de gaz à effet de serre. Ces gaz s'accumulent dans l'atmosphère, ce qui entraîne l'intensification de l'effet de serre et par conséquent le réchauffement de la planète.

Parmi les nombreuses conséquences établies, nous allons en aborder rapidement trois d'entre elles.

En premier lieu, la montée des températures : Une montée de plus de 0.8 degrés Celsius, des périodes caniculaires plus fréquentes, et les températures extrêmes qui touchent maintenant 10% de la planète contre 1% au 19^{ème} siècle.

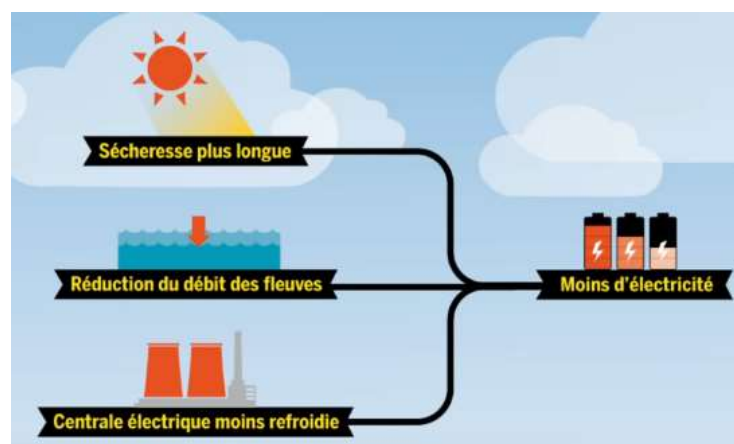


Figure 5: Centrales qui sont refroidies par les cours d'eau, répercussion sur la production d'électricité.⁸

De plus, la fonte des glaces contribue à élever le niveau de la mer. Si la banquise antarctique voit sa surface augmenter légèrement du fait d'un renforcement des vents, la

⁸ Le Monde. Comprendre le réchauffement climatique en 4 minutes. 3 :53 min. 2015. <https://www.youtube.com/watch?v=T4LVXCCmIKA>

banquise arctique a vu sa superficie fondre de 30% depuis les années 80. Le taux moyen d'élévation de l'océan est aujourd'hui de 3.3 mm/an, mais ce taux pourrait croître si les émissions se poursuivent. Dans le pire des scénarios, les océans pourraient s'élever d'1 mètre d'ici 2100, et plusieurs petits états insulaires du pacifique seraient amenés à disparaître.⁸

Enfin, un effet moins connu du public, mais un des plus inquiétants du changement climatique, reste l'acidification des océans. Le Co2 excédentaire se dissout en effet dans les eaux de surface et les rend plus acides. La rapidité de ce phénomène est inédite depuis 300 millions d'années et les scientifiques ignorent les effets qu'il y aura d'ici quelques décennies.⁸

Les coquillages et certains planctons qui constituent le socle de la chaîne alimentaire pourraient ne pas pouvoir s'y adapter. Leur disparition aurait un impact direct sur l'ensemble des espèces dont l'homme.

Après avoir fait un bref rappel sur l'impact du gaz à effet de serre sur l'environnement, penchons-nous sur l'impact du big data sur l'environnement plus spécifiquement. Nous pouvons alors nous demander si le Big data est polluant ? Mais auparavant, définissons exactement la pollution.

La pollution est l'association d'un polluant et d'un contaminant. C'est une substance synthétique ou naturelle que l'homme introduit dans l'environnement, ou dont l'homme modifie la répartition dans les différents compartiments de la biosphère. Un polluant est associé à un effet néfaste sur des êtres vivants, par exemple les algues vertes dans la manche. Un contaminant lui, signifie une présence inhabituelle dans un lieu, sans connaissance des effets potentiels.

Après cette précision, interrogeons-nous donc si le Big Data peut être considéré comme une source mondiale de pollution. Les éléments importants du Big Data peuvent se répartir en trois catégories : les équipements terminaux, les réseaux et les data centers qui consomment chacun une puissance électrique importante.

Représentant 2 % des émissions de gaz à effet de serre totales, l'empreinte carbone des data centers est comparable à celle du transport aérien. Le stockage des données est susceptible

de générer des émissions à hauteur de 14 % à l'horizon 2040. Dans une démarche pour un développement à grande échelle, la mise en place de data centers contribuerait donc rapidement au réchauffement climatique.⁹

La plupart des entreprises technologiques essaient maintenant de concevoir de nouveaux data centers qui seraient écologiquement viable et plus respectueux de l'environnement. Le meilleur moyen permettant d'éviter le risque de surchauffe des serveurs consiste à utiliser des systèmes de refroidissement alimenté par la même énergie utilisée par le data center.

C'est pourquoi, en 2009, Google a décidé d'ouvrir un data center à Hamina, en Finlande. Dans ce centre, les serveurs utilisent de l'eau de mer glaciale pour se refroidir. L'eau de pluie est également utilisée à cette fin. De plus, de son côté, Microsoft a mis en place un data center sous-marin. Ces techniques ont pour but de réduire l'impact des data centers sur le réchauffement climatique en limitant l'utilisation abusive de la climatisation.¹⁰



Figure 6: Data center de Google en Finlande¹¹

⁹ Joarson.2021. Data center : l'impact des infrastructures sur l'environnement et les solutions possibles.Lebigdata.fr

¹⁰ Jean-Marc Pierson et Laurent Lefèvre. 2015. Le Big Data est-il polluant ? CNRS- Le Journal

¹¹ Bastien L. Google investit 600 Millions dans son Data Center européen en Finlande

Selon des chiffres établis par une étude GreenIT publiée en octobre 2019, les 34 milliards de smartphones, ordinateurs, consoles de jeux et téléviseurs que compte la planète occupent une place centrale dans l'impact sur l'environnement.¹²

La construction d'outils informatiques impacte la quantité de ressources naturelles disponible. Cependant les ressources premières utilisées ne se renouvellent pas aussi rapidement que nous les consommons. Tous les ans, Global Footprint Network calcule avec WWF « le jour du dépassement », c'est-à-dire la date à partir de laquelle l'ensemble des ressources que la terre est capable de produire en un an a été consommé. En 2020, le jour du dépassement a été le 22 août, nous vivons donc 4 mois au-delà de nos capacités. Cette année, la date fatidique est tombée le 29 juillet, pour subvenir au besoin de la population mondiale de façon durable, il nous faudrait 1.7 Terre.¹³

En plus de la pollution générée par les data centers et par les équipements, celle produite par les utilisateurs est également à prendre en considération.

Le mail est une façon simple et rapide de communiquer et de plus en plus généralisé, pourtant son utilisation n'est pas sans impact sur l'environnement. S'il est accompagné d'une pièce jointe d'1 Mo, un mail émet 19 g de Co2, sa consommation électrique est équivalente à celle d'une ampoule allumée pendant une heure. ¹²

Cet impact pris individuellement est faible, mais quand on sait que 34 millions de mails sont envoyés toutes les heures sans compter les spams, les conséquences deviennent importantes. Cela correspondrait à l'équivalent de 14 tonnes de pétrole. Envoyer 20 mails par jour pollue autant que parcourir 100 km en voiture. ¹²

Certaines mesures sont possibles pour limiter l'impact, comme supprimer ces mails qui ne sont pas nécessaires. En effet un message conservé dans la boîte mail fait tourner les serveurs,

¹² 2020. L'incroyable impact de la pollution numérique et les bonnes pratiques à adopter très vite. Grizzlead

¹³ Inconnu.2021. 29 Juillet 2021, jour du dépassement de la Terre. Gouvernement

qui scannent nos emails en permanence. Compresser les pièces jointes et envoyer les photos en basse résolution. Ou bien, limiter le nombre de destinataires lors de l'envoi d'un email.

Il faut se rappeler que l'environnement de travail est très avide d'électricité, et d'équipements. De plus, l'impact du trajet domicile-travail en voiture pour les équipes IT représente une conséquence non-négligeable.

Mais aussi, la consommation des vidéos en ligne peut sembler être une activité sans répercussion, or elle consomme une quantité d'énergie phénoménale. L'une des raisons principales est le poids des fichiers consultés, qui nécessite plus d'énergie pour stocker ce fichier sur des serveurs. Le streaming vidéo correspond à 60% des flux de données sur Internet. Regarder une heure de vidéo consomme autant d'électricité qu'un réfrigérateur pendant une année. Une bonne pratique serait d'arrêter la lecture automatique des vidéos, le coût d'émission en Co2 réduirait alors.¹²

Pour finir, les objets connectés génèrent à eux seuls 39% des émissions de gaz à effet de serre du domaine du numérique. Ils contribuent également à hauteur de 76% à l'épuisement des ressources naturelles non-renouvelables du monde. On observe que 3 Français sur 4 possèdent un smartphone, mais seuls 6% d'entre eux sont recyclés dans le pays.¹²

La tendance qui se dessine n'est pas très optimiste, la planète comptera 48 milliards d'objets connectés en 2025 avec une empreinte environnementale 3 fois plus importantes qu'en 2010.

Même si le Big Data a un impact négatif sur l'environnement de par l'utilisation croissante d'ordinateurs, de smartphones et la construction de data centers, certaines initiatives montrent que le Big Data présente également un impact environnemental positif en utilisant à bon escient les données produites.

Ces nouvelles technologies deviennent progressivement indispensables à la protection de l'environnement. De l'encadrement de la pêche à la surveillance des zones protégées ou de l'étalement urbain, nombreuses sont leurs applications.

Comprendre l'état des milieux naturels et non naturels ainsi que leur évolution est vital pour la protection de l'environnement. Obtenir des informations précises de ces espaces est nécessaire et implique d'utiliser les moyens d'observation les plus efficaces.

Par exemple, en utilisant les observations satellitaires qui permettent d'obtenir des vues à grande échelle. Cette technique est particulièrement utile pour lutter contre l'artificialisation des sols et protéger la biodiversité. Mais lorsque des images plus précises sont nécessaires, il existe d'autres méthodes d'observation, tels que les drones civils.¹⁴

En raison de la diversité des sources d'information et de la complexité des phénomènes observés, les traitements de signaux et d'images posent de réels défis, car ceux-ci nécessitent des compétences techniques avancées. C'est pourquoi la combinaison des techniques de traitement des données avec la modélisation des phénomènes observés permettent d'obtenir des informations précises.

Le site Global Forest Watch, qui recense en temps réel la déforestation sur l'ensemble du globe, illustre bien le traitement des données issues des différentes techniques d'observation. Mentionnons aussi l'existence d'Ecosia, qui est un méta moteur de recherche et reverse les 80% de ses bénéfices générés par la publicité à l'aide pour la reforestation. En utilisant cet outil, les utilisateurs contribuent à réduire l'impact écologique à leur échelle.¹⁵

Comme nous le savons, l'utilisation de la technologie dernier cri pour la protection de l'environnement, bien qu'elle permette l'optimisation des réponses, présente aussi un coût environnemental.

¹⁴ Pierre Emmanuel Bouchez. 2014. Les nouvelles technologies au service de l'environnement ? Actu Environnement

¹⁵ Anais Perrenoud. 2019. Le Big data : Menace ou Solution pour l'environnement ? Carnet d'Idées - blog du Master IS

Etat de l'art

I_ La collecte des données

A) Vision générale

Nous avons vu ci-dessus quel est l'impact des nouvelles technologies sur le changement climatique. Nous sommes donc amenés à se demander, si l'utilisation de ces technologies afin de protéger l'environnement, ne peut-elle pas être considérée comme paradoxale ?

Nous allons présenter dans cette première partie une vue d'ensemble des applications du big data dans les études liées au changement climatique. Nous avons remarqué que les progrès informatiques, l'augmentation des ensembles de données écologiques et environnementales, ainsi que les différentes méthodes de collecte de données, révolutionnent le secteur de l'écologie. C'est pourquoi nous allons voir sur ce schéma quelles sont les multiples applications et créations de valeurs qu'engendre le Big data.

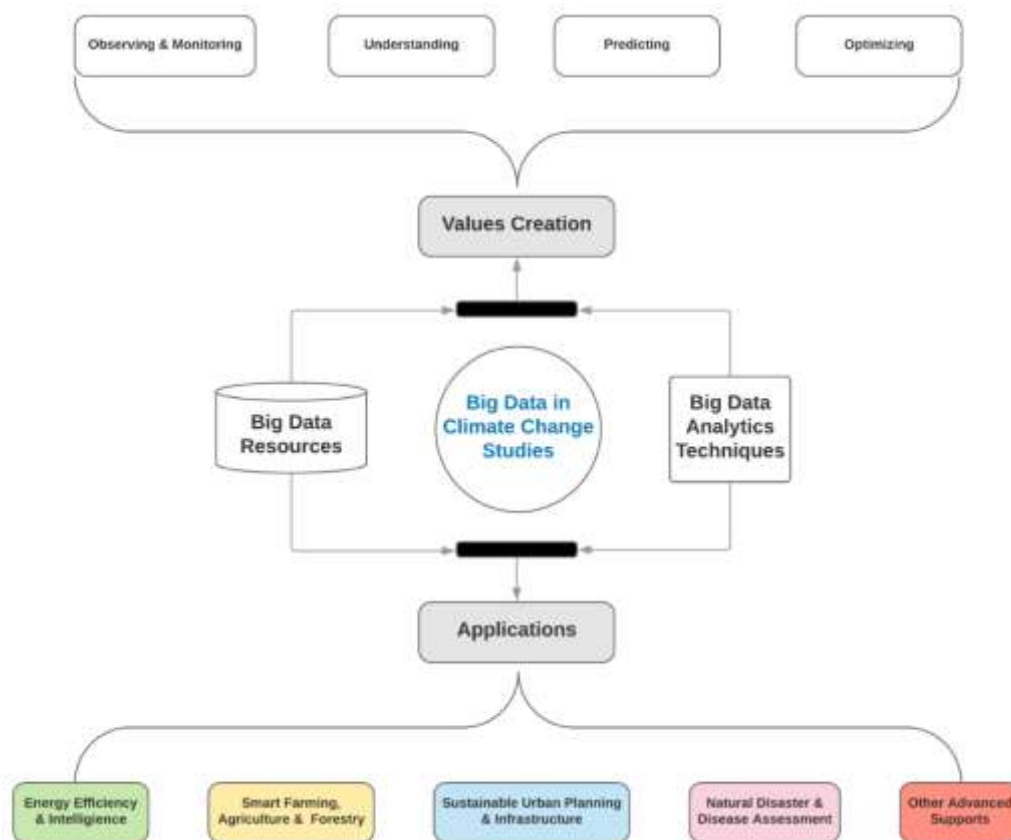


Figure 7: Cadre du big data dans les études sur le changement climatique. (Hossein Hassani 2019)

Comme le montre la figure 5, nous constatons que les applications du big data dans le domaine du changement climatique comportent deux éléments fondamentaux :

- Les ressources de big data climatiques
- Les techniques d'analyse du big data

Nous classons ces études en fonction de la création de valeur ainsi que du sujet spécifique de leur application. Afin de faciliter l'accès aux applications et de fournir des orientations claires, nous résumons que le big data dans le domaine du changement climatique fonctionne principalement dans quatre aspects de la création de valeurs : (Hossein Hassani 2019)

- l'observation et la surveillance
- la compréhension
- la prévision
- l'optimisation

Tandis que les applications sont regroupées en cinq thèmes :

- l'efficacité et l'intelligence énergétiques
- l'agriculture et la foresterie intelligente
- l'urbanisme et les infrastructures durables
- l'évaluation des catastrophes naturelles et des maladies
- d'autres supports avancés

Afin d'obtenir une observation approfondie du changement climatique, la technologie d'observation de la terre a joué un rôle important au cours des dernières décennies (Hua-Dong Guo 2015). Un système de big data multidimensionnel a été établi et continue de se développer rapidement, ce qui nous a permis d'observer et de surveiller les changements à l'échelle mondiale de divers paramètres terrestres et climatiques. Selon (Hossein Hassani 2019), les données climatiques proviennent généralement de quatre sources différentes :

- in situ (sur place)
- télédétection (C'est la discipline scientifique qui regroupe l'ensemble des

connaissances et des techniques utilisées pour l'observation, l'analyse, l'interprétation et la gestion de l'environnement à partir de mesures et d'images obtenues à l'aide de plates-formes aéroportées, spatiales, terrestres ou maritimes.)¹⁶

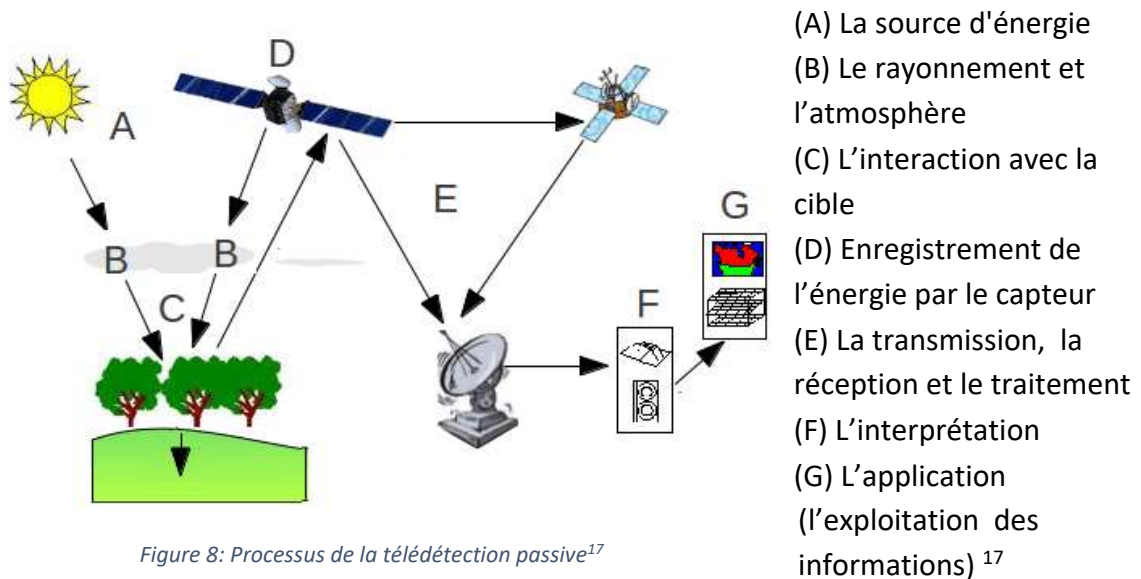


Figure 8: Processus de la télédétection passive¹⁷

- Les statistiques de sortie du modèle (MOS) (une technique de régression linéaire multiple dans laquelle prédictions, souvent des quantités proches de la surface, telles que la température de l'air, la direction, vitesse du vent).
- Paléoclimatique (le climat d'une ancienne époque géologique, il y a des milliers, voire des millions d'années).

¹⁶ Inconnu. 2019. Qu'est-ce que la télédétection. Université de Sherbrook

¹⁷ Chaima Grimene.2019. Classification de l'occupation des terres agricoles par télédétection satellitaire. researchgate.net

Le big data a également favorisé la progression rapide de l'analyse intensive des données dans les études liées au changement climatique.

Les abondantes ressources nous ont permis d'acquérir des connaissances sur ce que sont les grandes données climatiques, comment elles sont collectées et à quoi elles peuvent servir. Cependant, elles ne font qu'effleurer la surface de l'analyse des big data. (Hossein Hassani 2019)

B) Discussion

Nous allons aborder dans cette partie la collecte des données et son enjeu sur l'environnement au travers d'une application de collecte de données de biodiversité.

Tout d'abord l'E-recherche qu'est-ce-que c'est ? Nous pouvons l'expliquer comme l'application de différentes technologies de l'information et de la communication (TIC) pour une investigation scientifique. Elle utilise les réseaux informatiques nationaux, les laboratoires virtuels et le cloud pour le suivi et la collecte de données. Elle a le potentiel de faciliter la recherche et la connaissance scientifique dans le domaine de l'écologie de par sa capacité croissante à stocker, gérer, partager, analyser et synthétiser une grande quantité de données.

C'est grâce à cette combinaison entre le Big Data et l'E-recherche que la création du Laboratoire Virtuel sur la biodiversité et le changement climatique a pu voir le jour. Son but est d'aider la communauté de chercheurs en les connectant aux installations de recherche existante, aux ensembles de données, aux outils de calcul, d'analyse et de visualisation à haute définition.

La BCCVL (Biodiversity and Climate Change Virtual Laboratory) est actuellement alimentée par des jeux de données principalement autrichiens, mais la fourniture de jeux de données mondiaux et de liens vers des bases de données mondiales est en cours de développement. Cependant, la BCCVL peut actuellement être utilisée pour la modélisation de la distribution des espèces et d'autres analyses de la biodiversité partout dans le monde grâce à l'assimilation des données fournies par les utilisateurs. (Willow Hallgren 2016)

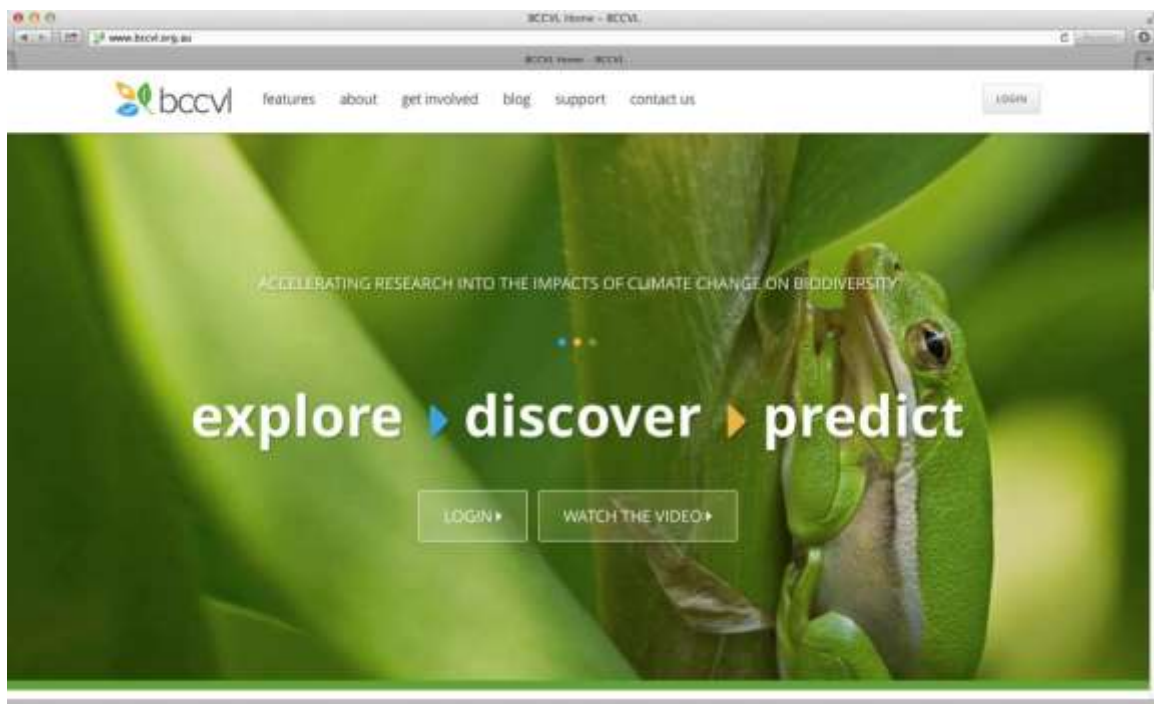


Figure 9: Présentation du Laboratoire virtuel BCCVL (Willow Hallgren 2016)

Elle permet aux chercheurs de mener des expériences de modélisation et des analyses connexes de manière beaucoup plus efficiente et efficace. Elle diminue le temps de préparation associé à la modélisation, y compris la préparation des données (c'est-à-dire l'identification, l'acquisition, la mise à l'échelle, la normalisation, la validation et la visualisation des données), la mise en place de l'environnement de modélisation, qui peut nécessiter l'apprentissage d'un langage de programmation comme, par exemple, le langage R, et l'écriture de scripts pour exécuter des expériences d'ensemble complexes.

La section "Datasets" du BCCVL contient des données sur la localisation et les caractéristiques des espèces, des données sur le climat actuel et futur, et d'autres données environnementales (par exemple, le sol, la géologie et le type de végétation).

L'architecture du BCCVL est conçue pour gérer de grands ensembles de données, traiter les données par le biais d'expériences, afficher les résultats des expériences et partager les données en toute sécurité dans un environnement cloud.

Le BCCVL est novateur dans son utilisation des technologies basées sur l'informatique dématérialisée pour réaliser des fonctions de modélisation traditionnellement réservées aux services de clusters ou aux systèmes de calcul de haute performance . Les technologies basées sur l'informatique dématérialisée sont souvent conçues à l'aide de matériel de base (calcul, stockage et mise en réseau) et permettent d'atteindre l'échelle et la résilience grâce à la possibilité d'ajouter et de remplacer facilement des composants individuels. (Willow Hallgren 2016)

Dans cette partie, nous allons constater que malgré les multiples progrès technologiques et techniques, les collectes de données sont diverses et orientées vers des objectifs de surveillance différents. De ce fait, nous examinerons comment l'essor de la détection citoyenne, ou l'utilisation de technologies numériques peu coûteuses et accessibles pour surveiller les environnements, a contribué à l'apparition de nouveaux types de données et de pratiques en matière de données environnementales.

En discutant de la recherche participative sur la détection de la pollution atmosphérique avec des habitants du nord-est de la Pennsylvanie préoccupés par les effets de la fracturation hydraulique, nous examinerons comment les nouvelles technologies de production de données environnementales posent également de nouveaux problèmes pour analyser et donner un sens aux données recueillies par les citoyens. Mais aussi comment les données environnementales soulèvent différentes préoccupations et possibilités par rapport aux Big Data, qui peuvent être distinctes des études sur la sécurité ou les médias sociaux.

Le terme « données citoyennes » a 2 significations :

- Les données générées par la population et les données recueillies en dehors du domaine scientifique en utilisant plusieurs techniques et technologies de surveillance.

- Ce que les citoyens peuvent déduire et formuler des différentes données collectées.

Les données citoyennes contribuent aux multiples flux de données qui peuvent donner naissance au Big Data et à ses techniques d'analyse émergentes. (Jennifer Gabrys 2016)

Les données « juste assez bonnes » sont les données générées par des instruments à faible technologie et à faible coût, ainsi que les données qui pourraient être observationnelles ou expérientielles.

Les données citoyennes pourraient initier des conversations avec les régulateurs environnementaux, pour affirmer ou non sur les processus polluants ou pour plaider en faveur de l'investissement de plus de ressources dans les infrastructures de surveillance.

En effet, le nombre croissant de capteurs numériques à faible coût est en train de révolutionner la science de la qualité de l'air. Les technologues et les régulateurs environnementaux suggèrent que les capteurs environnementaux puissent fournir des données environnementales là où il y a une absence de réseaux et d'infrastructures de surveillance.

Cependant, ces données n'ont pas nécessairement le même niveau de précision numérique ou d'exactitude que les données produites pour la conformité réglementaire. Néanmoins, leurs qualités permettent d'exprimer des tendances sur le fait que la pollution de l'air se produit à certains moments de la journée, dans des lieux communs ou distincts, et en relation avec les sources d'émissions, par exemple.

Alors que certaines agences comme l'EPA soulignent les possibles bénéfices de la surveillance citoyenne, un tel changement nécessiterait également une transformation dans les réseaux et les pratiques de surveillance pour être en mesure d'accueillir les données. (Jennifer Gabrys 2016)

Voyons l'exemple d'utilisation de données citoyennes. Dans la région du nord-est de la Pennsylvanie rurale, il y a un manque relatif d'infrastructure de surveillance de la qualité de l'air, puisque la plupart des contrôles de la qualité de l'air axés sur la réglementation sont

entrepris là où il y a des densités de population plus élevées et sont concentrés dans les zones urbaines. C'est pourquoi les habitants de cette région ont décidé de récolter des preuves de la pollution présente dans l'air et des dangers que leur santé encourt. (Jennifer Gabrys 2016)



Figure 10: Surveillance de la qualité de l'air par les citoyens du canton de Liberty (Jennifer Gabrys 2016)

L'ensemble de données récolté a fourni une nouvelle indication de ce qui pourrait être dans l'air, puisque les rapports individuels d'effets de santé ne comptaient pas comme une preuve suffisante ou concluante de problèmes de pollution ou d'exposition.

Les données recueillies par les citoyens sont uniques dans le sens où elles fournissent non seulement des données sur la qualité de l'air dans une région où la surveillance est relativement absente, mais aussi parce qu'elles sont denses dans l'espace et s'étendent sur une période de surveillance plus longue que la plupart des surveillances " expertes " épisodiques. Cela permet de comparer les données voisines et de corroborer les données des capteurs en les complétant par des observations.

Bien que de nombreux experts considèrent les normes réglementaires comme le principal, voire le seul moyen d'entreprendre une surveillance, les données générées par les citoyens et l'analyse des données peuvent offrir des façons alternatives de penser face aux discours souvent monolithiques autour des données environnementales et du Big Data. (Jennifer Gabrys 2016)

II_ L'analyse des données

A) Vision générale

Nous allons aborder dans cette partie les techniques courantes pour le big data dans les études et analyse des données sur le changement climatique, indépendamment des sujets d'application. L'un des principaux rôles de l'analyse des big data dans le domaine du changement climatique est la gestion et l'utilisation des ressources pour lutter contre le réchauffement de la planète. Il existe de nombreuses applications axées sur l'amélioration de l'efficacité énergétique et la mise en place de réseaux énergétiques intelligents. (Hossein Hassani 2019)

Les techniques sous-jacentes de l'analyse des big data sont principalement regroupées en quatre catégories : le regroupement, la classification, la régression et la fouille de règles d'association.

De plus, (Witten, et al. 2016) résume également que les dix algorithmes d'exploration de données les plus influents sont C4.5, k-mean, SVM (support vector machine), Apriori, EM (expectation maximization), PageRank, AdaBoost, kNN (k-nearest neighbors), Navie Bayes et les arbres de classification et de régression (CART), pour lesquels ils ont été incorporés dans presque tous les systèmes d'analyse de big data commerciaux et open source. (Hossein Hassani 2019)

Il est également important de préciser que les progrès de ces techniques et de nouveaux algorithmes se développent quotidiennement, parallèlement aux progrès technologiques d'aujourd'hui.

Sur la figure 11 ci-dessous, nous pouvons voir les différentes techniques utilisées pour l'analyse de big data, nous allons les détailler par la suite.

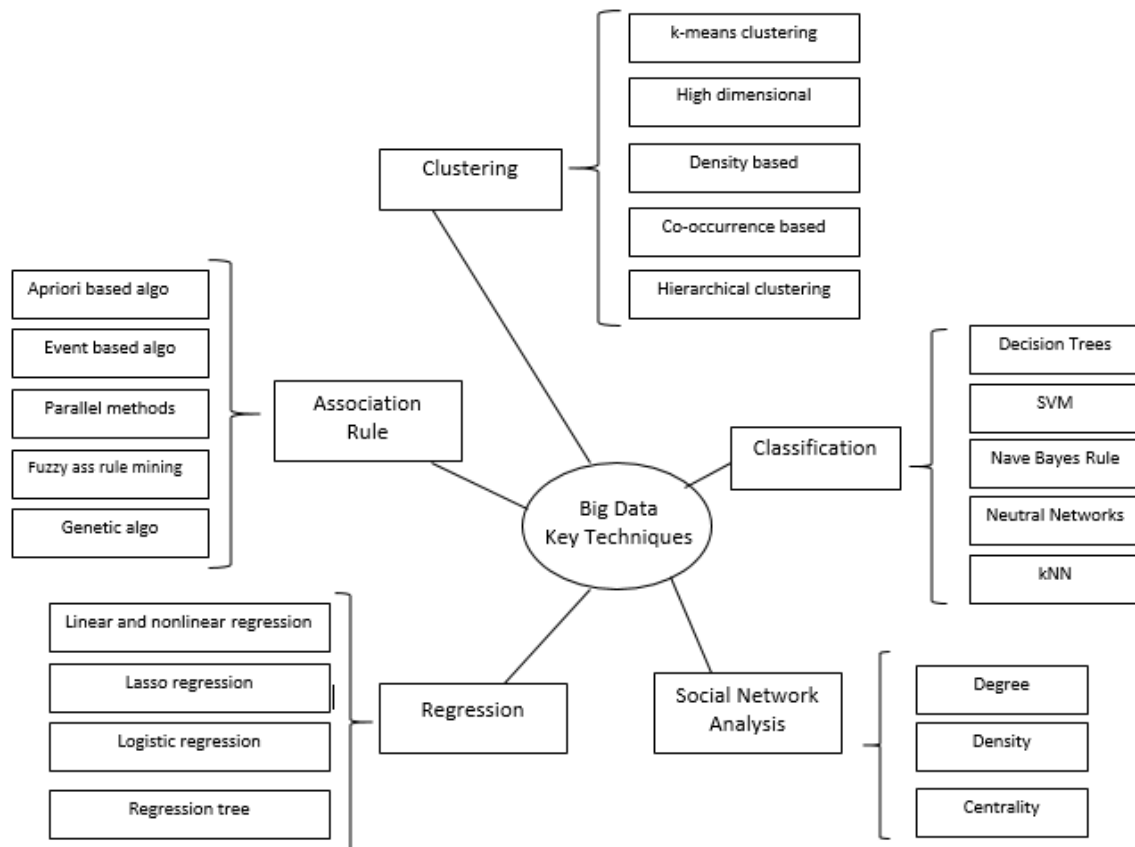


Figure 11: Structure des techniques clés du Big data. (Hossein Hassani 2019)

Le **Clustering** a pour objectif principal de regrouper des objets de données similaires ou étroitement liés par le biais de l'exploration des données, afin de séparer les objets différents. Il comprend une variété d'approches que l'on peut voir sur le schéma ci-dessus.

La **Classification**, quant à elle, permet de classer les objets de données dans des groupes prédéfinis. Il s'agit de l'une des techniques d'analyse de Big data les plus fondamentales, avec un ensemble de méthodes bien établies.

L'une de ses méthodes les plus connues est « *Decision tree* », qui applique une série de questions élaborées afin de réaliser les tâches de classification des attributs.

Les « *Support Vector Machines (SVM)* », elles, divisent les objets en deux classes en utilisant un hyperplan de séparation optimal conçu pour minimiser l'erreur de classification.

Un autre classificateur, appelé la « *Naive Bayes Rule* », applique la règle de Bayes pour calculer la probabilité de chaque classe. Cette méthode est basée sur l'hypothèse que tous les attributs sont indépendants, et la classe de prédiction peut être déterminée pour celle dont la probabilité postérieure est la plus élevée. Pour finir, l'approche « *kNN* » permet d'identifier les plus proches voisins de l'objet observé afin de déterminer son étiquette de classe.

Association Rule Mining a été proposée pour la première fois sur l'ensemble des données de supermarché et visait à étudier les cooccurrences entre les objets de données. Il s'agit d'une technique permettant d'identifier les occurrences simultanées qui se produisent plus fréquemment que la fréquence moyenne de cooccurrence existant dans l'ensemble de données.

Regression est également considérée comme une importante technique d'analyse des données volumineuses en raison de ses capacités de réduction des dimensions, d'extraction d'informations, d'estimation et de prédiction. Le concept fondamental de la régression est d'étudier la relation entre deux ou plusieurs variables afin de faciliter la prévision et la prise de décision. Il existe une variété de techniques de régression bien établies comme la régression linéaire et non linéaire, la régression lasso, la régression logistique et l'arbre de régression, etc.

Social Network Analysis (SNA) repose sur le principe de la théorie des graphes et constitue une technique relativement nouvelle d'analyse des données massives. Elle étudie les connexions et le contenu entre les objets dans une pile massive d'informations afin de structurer un réseau social. Le réseau social de base est formé de nœuds et de nœuds apparentés reliés par des liens.

Il est à noter que l'application dans certains domaines de l'analyse des Big data au changement climatique semble déséquilibrée. Les raisons possibles sont, un manque de ressources de données correspondantes, un manque de profit économique et de financement de la

recherche, une mauvaise communication entre des groupes de chercheurs aux compétences différentes ou bien des informations incertaines. (Hossein Hassani 2019)

On peut donc se demander par la même occasion, pourquoi l'impact de la science des données sur la science du climat est-il plus lent que dans d'autres domaines ? Nous avons identifié trois facteurs qui ont participé au ralentissement des progrès.

Nous allons faire un petit aparté pour identifier le but de la science du climat. Elle est l'étude de l'environnement de notre planète. Fondamentalement, elle se concentre sur l'étude des changements à grande échelle dans les terres, l'atmosphère, les océans et la cryosphère sur de longues périodes, bien que certains considèrent que des échelles de temps plus courtes comme les jours ou les semaines font également partie de l'étude du climat.

Revenons aux facteurs qui participent au ralentissement des progrès. Tout d'abord, les données utilisées par la science du climat vont à l'encontre de bon nombre des hypothèses et des pratiques de la science des données traditionnelles. Par exemple, la majorité des données climatiques sont organisées selon une grille spatio-temporelle. En tant que telles, les données sont auto-corrélées, les régions proches dans l'espace ou dans le temps ayant tendance à être fortement corrélées. Par conséquent, toute méthode qui impose des hypothèses d'indépendance entre les points de données n'aura qu'un intérêt pratique avec de telles données.

Deuxièmement, le domaine de la science des données s'est par le passé concentré spécialement sur certaines tâches et mesures d'évaluation qui ne sont pas applicables à certains besoins de la science du climat.

Enfin, la science du climat, ses données et ses défis n'ont été exposés à la communauté plus large de la science des données que récemment. (James H. Faghmous 2014)

TABLE 1. LIST OF CLIMATE DATA SOURCES
WITH THE TEMPORAL COVERAGE OF EACH DATA SOURCE

	<i>Key strength</i>	<i>Key weakness</i>
Climate model	Ability to run forward simulations	Relies solely on physics
<i>In situ</i> observation	Direct observations	Spatial bias
Satellite	Global coverage	Lack of continuity as missions last on average 5 years
Paleoclimate	Ability to use proxy data to infer preindustrial climate trends	Techniques to analyze such data are still evolving

Figure 12: Liste des sources de données climatiques (James H. Faghmous 2014)

Évoquons maintenant, les raisons pour lesquelles il faut comprendre d'où viennent les données. Premièrement, il est nécessaire de comprendre comment les données sont générées, quel est leur objectif et quels sont les processus de génération, cela permet d'avoir un fil conducteur. Deuxièmement, comprendre les biais appartenant aux données nous donne une chance de se corriger ou d'ajuster les résultats.

Enfin, il est nécessaire de développer de nouvelles méthodes d'analyse. La majorité des analyses de réseau dans le domaine du climat reposent sur des caractéristiques de réseau dérivées des mathématiques et d'autres domaines, qui peuvent ne pas convenir à l'application en question.

Un grand avantage de la science des données traditionnelles est la définition claire des tâches d'apprentissage (régression, classification, etc.). Dans la science du climat, cependant, la fonction objective peut être plus difficile à définir.

Par exemple, l'un des impacts les plus redoutés du changement climatique mondial est la sécheresse. Or, la notion même de sécheresse est ambiguë du point de vue des données. Premièrement, il existe de nombreux types de sécheresses : agricoles, météorologiques et hydrologiques. Cependant, en général, les sécheresses peuvent être définies comme " l'absence prolongée ou l'insuffisance marquée de précipitations ". Deuxièmement, la façon de

représenter cette " insuffisance " n'est pas claire, par exemple, les sécheresses peuvent être quantifiées en termes absolus ou relatifs, et selon cette représentation, les conclusions peuvent être différentes.¹⁸

Deux études récentes publiées dans Nature Climate Change et Nature sont parvenues à des conclusions opposées quant à l'existence de changements notables dans les tendances de la sécheresse en raison du changement climatique. Une étude a mis en évidence les résultats différents en soulignant que chaque étude s'appuyait sur différent ensemble de données de précipitations. Il existe d'autres cas où deux études ont utilisé les mêmes données, mais sont arrivées à des conclusions opposées. Car les deux groupes se sont appuyés sur des techniques d'analyse de données différentes. (James H. Faghmous 2014)

Plus largement, les deux risques les plus courants de l'analyse des Big data lorsqu'il s'agit de domaines scientifiques concernent les méthodes utilisées et les moyens par lesquels les résultats sont interprétés.

Du point de vue de la méthode, certaines des techniques d'analyse de données les plus populaires sont mal adaptées à l'analyse de données bruyantes, hétérogènes et auto corrélées que l'on trouve dans les sciences climatiques. Ainsi, bien que ces méthodes produisent certainement des résultats, elles ne permettent que rarement de les comprendre. Le plus grand risque est l'interprétation de ces résultats dans un état d'esprit sans théorie. Le chercheur en Big data doit être averti qu'il doit interpréter tout résultat dans le cadre des théories connues du domaine d'application. (James H. Faghmous 2014)

Par conséquent, l'analyse des big data ne doit pas être considérée comme la "solution miracle" de la recherche moderne et doit être utilisée en complément d'autres outils, de peur d'en faire un mauvais usage. Ce qu'il faut, c'est développer des méthodologies axées sur les données et guidées par la théorie pour limiter la recherche, découvrir des modèles plus significatifs et produire des modèles plus précis.

¹⁸ Inconnu. 2013. Climat et sécheresse. Météo France

B) Discussion

La science du climat est l'un des plus grands défis de notre époque en raison du danger qu'elle représente pour notre planète. Elle représente également une formidable opportunité pour la recherche sur les big data, car la science du climat fait apparaître des données, des méthodes et des défis d'évaluation inédits. Cependant, l'analyse des big data ne suffit pas à elle seule à explorer les données climatiques de manière perspicace et précise. Il est nécessaire de disposer de méthodes de science des données guidées par la théorie, qui associent la puissance de l'analyse des big data à la prudence de la théorie. (James H. Faghmous 2014)

Les techniques et algorithmes traditionnels de stockage de données ne sont pas applicables pour traiter l'énorme quantité de données climatiques. Nous allons aborder dans cette partie, un cadre de traitement des données évolutif avec un nouvel algorithme de détection des changements. Le grand volume de données climatiques est stocké sur le système de fichiers distribués Hadoop (HDFS) et l'algorithme MapReduce est appliqué pour calculer la moyenne saisonnière de divers paramètres climatiques tels que la température maximale, la température minimale, les précipitations, le vent, l'humidité relative et le rayonnement solaire. Les grandes données climatiques sont réduites à l'aide du cadre Hadoop MapReduce, de plus un algorithme de somme cumulative spatiale est proposé pour surveiller les changements saisonniers dans les données climatiques.

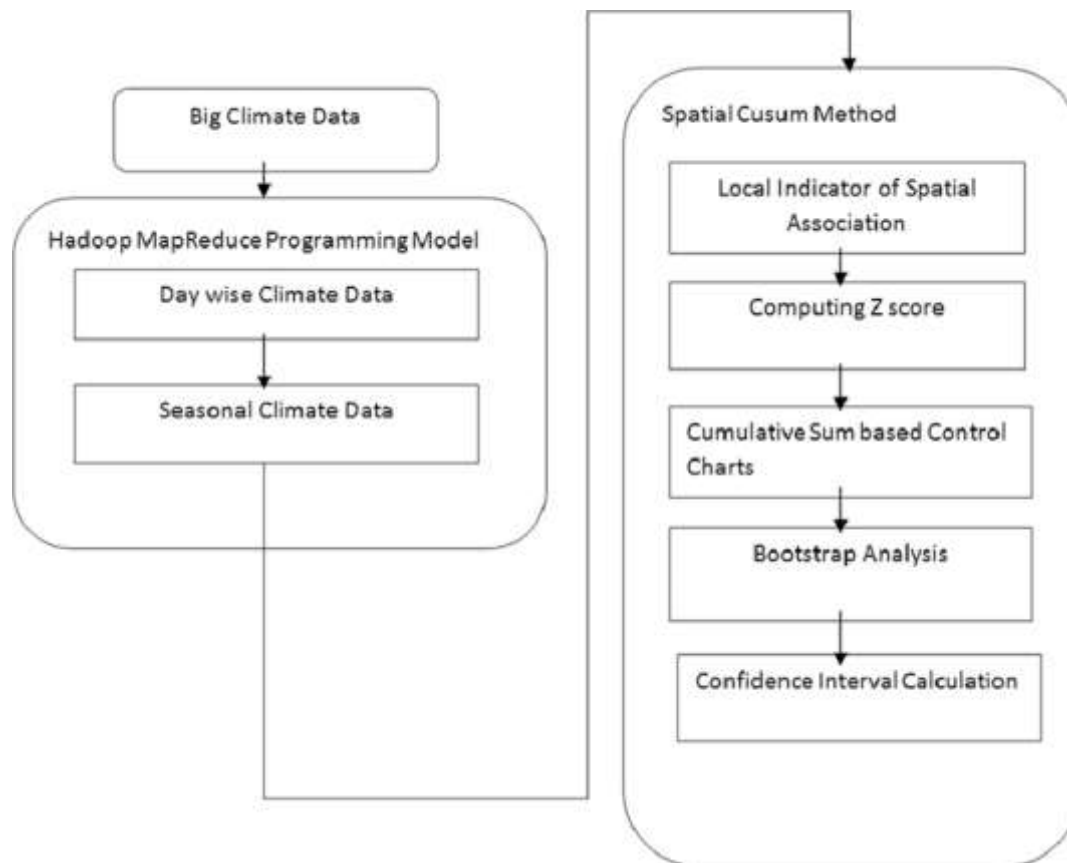


Figure 13: Cadre proposé pour la détection du changement climatique. (Gunasekaran Manogaran 2018)

Cette partie explique les différentes approches existantes au service de la détection du changement climatique, telles que la méthode de la somme cumulative et la méthode d'analyse bootstrap.

Méthode de la somme cumulée

Les méthodes de somme cumulative sont utilisées pour détecter les changements lents et drastiques. Cette méthode est utilisée dans de nombreuses applications, notamment pour surveiller les changements dans l'environnement de production, l'anticipation des maladies, la déforestation et l'analyse de la criminalité. (Gunasekaran Manogaran 2018) utilise la méthode de la somme cumulative pour surveiller les changements climatiques.

Taylor a développé la méthode d'analyse des points de changement à l'aide de graphiques de somme cumulative (CUSUM) et de méthodes d'analyse bootstrapping. La carte de contrôle CUSUM peut être calculée comme suit :

Calculez la moyenne pour "n" points de données $X_1, X_2, \dots, X_n$, par l'équation suivante : La valeur de la somme cumulée SI est calculée sur la base de l'équation suivante :

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

The cumulative sum value S_i is calculated based on the following equation :

$$S_i = S_{i-1} + (X_i - \bar{X}) \text{ for } i = 1, 2, \dots, n$$

Where,

$$S_0 = 0$$

Calculate the maximum and minimum S_{max} and S_{min} , by the following equation:

$$S_{max} = \max_{i=0,1,\dots,n} S_i$$

$$S_{min} = \min_{i=0,1,\dots,n} S_i$$

Calculate the S_{diff} Values to find the changes incumulative sum value SI, by the following equation:

$$S_{diff} = S_{max} - S_{min} \text{ (Gunasekaran Manogaran 2018)}$$

Méthode d'analyse Bootstrap

La méthode de la somme cumulée est utilisée pour trouver les décalages ou les changements notables dans la moyenne. Cependant, l'analyse bootstrap est une métrique supplémentaire utilisée pour déterminer les changements significatifs calculés en réordonnant de manière aléatoire les " n " valeurs initiales. Cette analyse bootstrap est utilisée pour vérifier les

changements initialement calculés à partir de la valeur de la somme cumulée S_I . Le niveau de confiance est calculé pour les résultats de l'analyse bootstrap. Un seul bootstrap comprend les étapes suivantes :

- **Étape 1** : Réorganiser les 'n' valeurs originales $X_1^0, X_2^0, \dots, X_n^0$ en utilisant la méthode d'échantillonnage sans remplacement.
- **Étape 2** : Calculer les valeurs de la somme cumulative $S_1^0, S_2^0, \dots, S_n^0$.
- **Étape 3** : Calculer les valeurs maximales, minimales et de différence S_{max}^0, S_{min}^0 and S_{diff}^0
- **Étape 4** : Identifier la différence bootstrap S_{diff}^0 est inférieure à la différence originale S_{diff} .

Le rôle vital de l'analyse bootstrap est d'identifier le niveau significatif des résultats de la somme cumulative si aucun changement n'a eu lieu. Un nombre d'analyses bootstrap est effectué pour calculer le niveau de confiance des changements.

Le niveau de confiance CI peut être calculé, par l'équation suivante :

$$Confidence Level (CI) = 100 * \frac{X}{N} \%$$

Where,

N = Number of bootstrap samples performed

X = Number of bootstraps for which $S_{diff}^0 < S_{diff}$

En général, 90%, ou 95% de niveau de confiance sont nécessaires pour identifier le changement important dans les données originales. En outre, il n'est pas possible d'effectuer une analyse bootstrap pour n! . Toutefois, une analyse bootstrap de 1000 est suffisante pour trouver les changements importants dans les données originales. (Gunasekaran Manogaran 2018)

III_ L'exploitation des données

A) Vision générale

Pour finir, après avoir étudié les différentes techniques du big data dans l'analyse des données sur le changement climatique, nous allons évoquer l'exploitation des bases de données pour évaluer les impacts du changement climatique. L'analyse de données permet de mettre en évidence des tendances des données ou corrélations entre les données non-évidentes. L'exploitation de ces données va permettre de tirer profit de son analyse afin d'en tirer des conclusions.

Un défi de l'adaptation au changement climatique est de pouvoir confronter des bases de données climatiques avec des données environnementales ou sociétales. Cependant, cette tâche rencontre une première barrière au langage de chaque discipline, qui utilise un jargon et des notions qui lui sont propres. Un travail de traduction est donc essentiel pour faire des liens entre la discipline et les bases de données qu'elles produisent.

Pour la science du climat, cette traduction peut s'effectuer en transformant des données brutes due à l'observation ou à la simulation, en données élaborées qui sont pertinentes pour une autre discipline.

La seconde barrière est fondamentale, car elle s'apparente aux grandeurs elles-mêmes. Identifier les variables d'une base de données (par exemple climatique), qui interagissent avec celles d'une autre base de données (par exemple économique) est un véritable défi. Des hypothèses sont formulées en examinant des cas synchrones dans plusieurs BDD. (Yiou 2013)

Nous allons aborder un projet qui montre l'utilité de confronter des bases de données climatiques obtenues par satellite avec des données environnementales ou sociétales obtenues par les médias. Dans le domaine de l'espace et de la sécurité, le principal défi est

d'améliorer la capacité à extraire au bon moment des informations utiles et claires d'une énorme quantité de données hétérogènes.

Le défi sociétal "Secure Societies" a été défini pour protéger la liberté et la sécurité de l'Europe et de ses citoyens. Les principaux objectifs de ce défi sont de renforcer la résilience de notre société face aux catastrophes naturelles et d'origine humaine, de développer de nouvelles solutions pour la protection des infrastructures critiques. Une activité majeure pour soutenir ces objectifs est la fourniture de produits et services géospatiaux, principalement issus de données satellitaires. (Albani 2016)

Le projet pilote "Secure Societies" mis en œuvre pour valider la plateforme BigDataEurope se concentre sur l'intégration et la fusion des données provenant de la télédétection et de la détection sociale afin d'ajouter de la valeur aux pratiques actuelles d'exploitation des données; cet aspect est essentiel dans le domaine de l'espace et de la sécurité, où des informations utiles peuvent être tirées non seulement des données satellitaires, mais aussi des données provenant des médias sociaux et d'autres sources. Les outils automatiques pour la gestion et le traitement des données sont l'un des aspects clés, où les solutions adoptées doivent être intégrées dans l'ensemble de la chaîne de données afin d'effectuer efficacement l'analyse des données.

Le projet pilote BigDataEurope couvre toutes les questions liées aux Big Data, à savoir le volume avec les images satellites de grandes tailles, la variété avec les données hétérogènes telles que des images satellites et du contenu textuel, la vitesse avec les données sociales et le flux d'informations rapides.

Le pilote prend en compte deux flux de données différents :

- Le flux de détection des changements qui prend des images satellites pour détecter les zones présentant des changements dans la couverture ou l'utilisation des sols à l'aide de techniques de détection des changements ; les zones d'intérêt identifiées sont ensuite associées aux médias sociaux et aux actualités et présentées à l'utilisateur final pour une validation croisée.

- La procédure inverse est appliquée aux flux de travail de détection d'événements. La détection d'événements est déclenchée par les nouvelles et les informations des médias sociaux, où les sujets avec une connotation géospatiale constituent un événement localisé dans le temps et l'espace ; si tel est le cas, les images satellites correspondantes sont acquises et traitées afin de vérifier les changements dans la couverture ou l'utilisation des terres. (Albani 2016)

Concernant l'architecture du projet pilote, il a été conçue pour s'adapter aux deux flux de travail et se compose des 3 composants décrits dans la *Figure 14: Architecture du pilote* l'interface utilisateur, qui est une version modifiée de l'application web Sextant ; le composant de stockage, qui implique Apache Cassandra4 et Strabon ; le composant central, qui se compose de deux modules, l'un pour la détection des changements dans l'utilisation du sol à partir d'images satellites et l'autre pour la détection des événements dans les articles d'actualité et les médias sociaux. Le premier composant s'exécute côté client et constitue une application web qui peut être déployé sur diverses plateformes. Les deux derniers composants fonctionnent côté serveur afin d'offrir une évolutivité et une efficacité élevées. (Albani 2016)

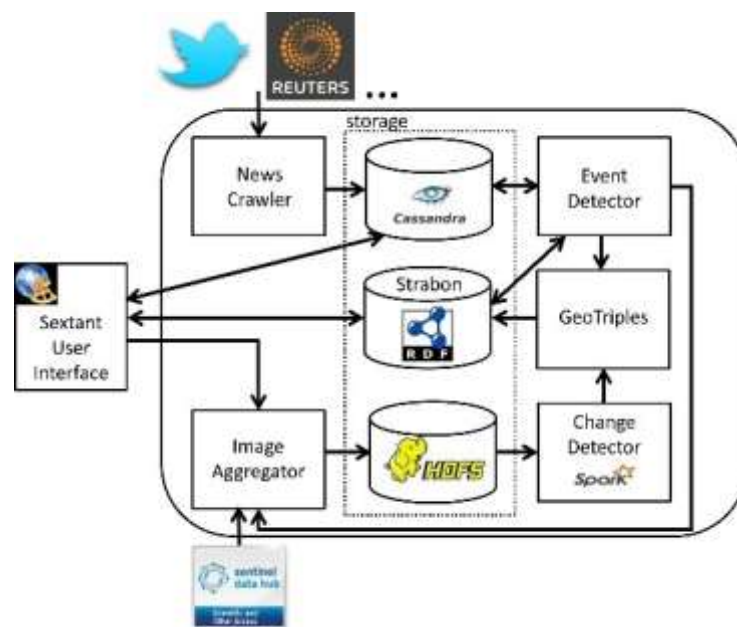


Figure 14: Architecture du pilote (Albani 2016)

Pour le flux de travail de détection des changements, Sextant offre une interface similaire à Google Earth, les coordonnées de la zone spécifiée sont extraites et transmises à la composante centrale du pilote. Pour la télédétection, les images GRD (Ground Range Detected) de Sentinel-1 ont été choisies pour cette première phase de mise en œuvre. Sentinel-1 bénéficiera, entre autres, de services liés à la surveillance de la surface terrestre pour les risques de mouvement, à la cartographie pour la gestion des forêts, de l'eau et des sols, et à la cartographie pour soutenir l'aide humanitaire et les situations de crise. (Albani 2016)

En ce qui concerne la détection sociale, l'accent est mis sur deux sources d'information complémentaires : les médias sociaux, représentés par Twitter, et les agences de presse. Dans le premier cas, il s'agit de texte brut (messages Twitter) accompagné de métadonnées au format JSON, tandis que dans le second cas, il s'agit d'articles d'actualité mis à disposition par des flux RSS au format XML. Ce pilote utilise l'API8 gratuite Twitter Public Streams, qui fournit un échantillon aléatoire de son contenu. Twitter propose de nombreuses options de paramétrage, ce qui permet au pilote de tirer le meilleur parti du contenu récupéré, même s'il s'agit d'un sous-ensemble des tweets générés.

Dans le cadre du pilote, le contenu est utilisé pour l'extraction de métadonnées, tandis que l'utilisateur est dirigé vers la source originale pour plus d'informations, afin de respecter les exigences en matière de droits de propriété intellectuelle. (Albani 2016)

B) Discussion

Évoquons maintenant un exemple du big data et son impact indirect sur le changement climatique. En effet, une exploitation ainsi qu'une analyse de données ont été menées par des chercheurs en communication environnementale qui se sont concentrés sur le rôle des cadres médiatiques dans la formation de l'opinion publique, en examinant le flux des conversations Twitter sur le changement climatique sur deux ans.

Ainsi, l'article (S. Mo Jang 2015) démontre que ce sont les cadres de canulars qui remettent en question la réalité du changement climatique. Ceux-ci sont prédominants aux États-Unis, en particulier dans les "États Républicains". Au contraire, au Royaume-Uni, Canada, Australie et dans les " États Démocrates " aux États-Unis ceux-ci sont moins présents.

Grâce à de nouveaux outils de collecte et d'analyse de données, il est désormais possible d'analyser et exploiter les conversations en ligne qui sont volontairement générées et partagées par les utilisateurs. Dans le présent article, (S. Mo Jang 2015) analyse les données de l'une des plateformes de réseaux sociaux les plus populaires (Twitter) sur deux ans et dans quatre pays anglophones. L'article va mettre en évidence les termes que les gens utilisent lorsqu'ils discutent du changement climatique, et les compare entre les pays et les États.

Des études précédentes suggèrent que les sceptiques du changement climatique ont tendance à utiliser "réchauffement global" au lieu de "changement climatique" pour discréditer l'authenticité des affirmations scientifiques. La recherche a indiqué que les différences au niveau national sont souvent dues à des pratiques différentes en matière de représentation des médias de masse qui encadrent les questions liées au changement climatique.

Pris dans leur ensemble, cinq thèmes communs apparaissent fréquemment dans les médias de masse, à savoir si le risque est présent, si l'affirmation scientifique est un mensonge, si le risque est causé par les activités humaines, les conséquences potentielles du risque et la façon de gérer le risque. (S. Mo Jang 2015)

Ces cinq cadres sont tous associés à la façon dont le public perçoit et traite un risque environnemental potentiel. En général, la couverture américaine se concentre sur la question de savoir si le changement climatique est un problème social légitime qui mérite l'attention du public alors que les médias des autres pays industrialisés se concentrent sur la manière de faire face à un défi environnemental connu.

Le changement climatique mondial est devenu un sujet très polarisé aux États-Unis, Le Pew Research Center for the People and the Press¹⁹ montre que seulement 39% des républicains, mais 85% des démocrates pensent que la protection de l'environnement devrait être considérée comme une priorité absolue. L'augmentation au fil des années, des personnes qui pensent que la lutte contre le changement climatique devrait être une priorité ne se fait que du côté des démocrates.

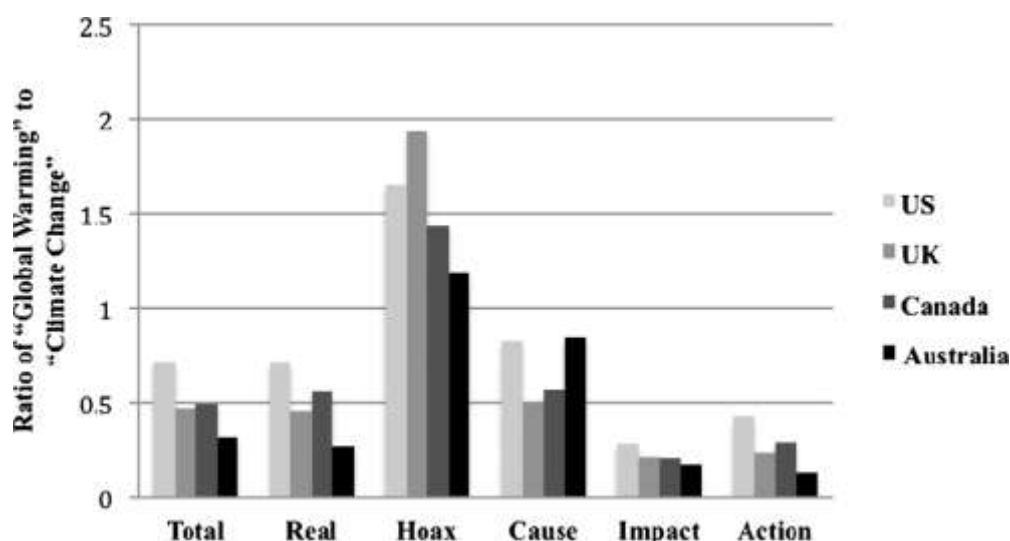


Figure 15: Ratio du « réchauffement de la planète » par rapport au « changement climatique » par cadres dans quatre pays. (S. Mo Jang 2015)

La collection, analyse et exploitation des grandes données pour limiter le changement climatique ne se limite pas qu'aux données climatiques. Comme le montre cet exemple les données médiatiques concernant l'opinion publique sont très importantes afin de pouvoir faire évoluer les mentalités sur la réalité du changement climatique.

¹⁹ Cary Funk and Brian Kennedy.2020. How Americans see climate change and the environment in 7 charts. Pew Research Center

Conclusion

Tout au long de cet état de l'art, nous avons recueilli différentes études et projets pour tenter de répondre à notre réflexion du départ : Quel est l'impact du big data sur le changement climatique et la pollution ?

Tout d'abord, nous avons étudié plusieurs exemples de projets comme « Baltimore Air » ou bien « Ecosia » qui illustrent les différentes solutions possibles pour que le big data ait un impact positif sur le changement climatique. Nous avons bien évidemment abordé aussi ses effets néfastes, comme la pollution que les data centers engendrent.

Ensuite, nous avons relevé différents projets qui sont menés à bien pour aider la collecte de grandes données, comme l'outil (BCCVL) qui recense la biodiversité pour aider les chercheurs du monde entier à centraliser leur données, mais aussi le projet de l'état de Pennsylvanie qui vise à démocratiser la collecte de donnée, tout en ayant l'encadrement nécessaire même si les techniques employées ne sont pas traditionnelles.

Par la suite, nous avons abordé les différentes techniques d'analyse de données comme la classification ou le clustering, nous avons aussi vu quelques méthodes statistique d'analyse comme la méthode de sommes cumulées ou la méthode d'analyse bootstrap qui permet de déterminer les changements significatifs de données.

Enfin, nous avons présenté un projet exploitant des bases données climatiques obtenues par satellite et les confrontant avec des données environnementales et sociétales obtenues par les médias. Dans le même thème, nous avons parlé d'une étude démontrant l'importance et l'influence des données venant des médias, concernant le changement climatique, sur l'opinion publique.

En résumé, comme nous l'avons constaté, le big data est un outil puissant, avec de larges perspectives dans les domaines économique et social. Compte tenu de la situation alarmante dans laquelle nous nous trouvons et des solutions qu'il peut apporter, il semble évident de

l'utiliser à bon escient pour l'environnement. Aujourd'hui, l'usage de la technologie est devenu quotidien et important. En revanche, l'impact de l'utilisation de ces données sur l'environnement reste encore flou.

D'ailleurs, l'entreprise Google est le parfait exemple de cet usage important des Big Data, par l'indexage de millions de documents afin de faciliter et d'accélérer la recherche. Pourtant, cela représente un coût environnemental désastreux. Les données sont analysées pour pouvoir proposer aux utilisateurs des contenus publicitaires, or pour réaliser ces traitements, nous estimons que Google possède plus d'un million de serveurs.

On peut se demander quelles conséquences auront toutes ces nouvelles technologies dans un futur proche ? Et si le big data n'aurait pas un plus un impact négatif que positif sur le changement climatique ? Mais aussi, à partir de quel volume de données consommées auront nous atteint le point de non-retour ?

Nous savons que les technologies de l'information sont de féroces consommatrices d'énergie, et produisent des gaz à effet de serre dans toutes les étapes de leur cycle de vie. C'est pourquoi, il nous appartient de veiller à ce que le coût environnemental de ces technologies soit contrebalancé, au moins en partie, par des progrès dans la lutte contre le réchauffement climatique et la pollution.¹⁰

Références

- Albani, S., M. Lazzarini, M. Koubarakis, Efi Karra Taniskidou, G. Papadakis, V. Karkaletsis and George Giannakopoulos. «A pilot for Big Data exploitation in the space and security domain.» 2016.
- Geetam S. Tomar, Narendra S. Chaudhari, Robin Singh Bhadoria, Ganesh Chandra Deka. *The Human Element of Big Data: Issues, Analytics, and Performance*. Édité par CRC Press. Taylor & Francis Group, 2016.
- Gunasekaran Manogaran, Daphne Lopez. «Spatial cumulative sum algorithm with big data analytics for climate change detection.» *Computers & Electrical Engineering* 65 (January 2018): 207-221.
- Hossein Hassani, Xu Huang, Emmanuel Silva. «Big Data and Climate Change.» *Big Data Cognitive Computing* 3, n° 1 (feb 2019): 12.
- Hua-Dong Guo, Li Zhang, Lan-Wei Zhu. «Earth observation big data for climate change research.» *Advances in Climate Change Research* 6, n° 2 (June 2015): 108-117.
- James H. Faghmous, Vipin Kumar. «A Big Data Guide to Understanding Climate Change: The Case for Theory-Guided Data Science.» *Big Data* 2, n° 3 (September 2014): 155-163.
- Jennifer Gabrys, Helen Pritchard, Benjamin Barratt. «Just good enough data: Figuring data citizenships through air pollution sensing and data stories.» *Big Data & Society* 3, n° 2 (December 2016).
- Lacey, E.A., Hammond, T.T., Walsh, R.E. et al. «Climate change, collections and the classroom: using big data to tackle big problems.» *Evolution: Education and Outreach* 10, n° 2 (2017).
- S. Mo Jang, P. Sol Hart. «Polarized frames on “climate change” and “global warming” across.» *Global Environmental Change* 32 (2015): 11-17.

Willow Hallgren, Linda Beaumont, Andrew Bowness, Lynda Chambers, Erin Graham, Hamish Holewa, Shawn Laffan, Brendan Mackey, Henry Nix, Jeff Price, Jeremy Vanderwal, Rachel Warren, Gerhard Weis. «The Biodiversity and Climate Change Virtual Laboratory: Where ecology meets big data.» *Environmental Modelling & Software* 76 (February 2016): 182-186.

Witten, I.H., E. Frank, M.A. Hall, et C.J. Pal. «Data Mining: Practical Machine Learning Tools and Techniques;.» *SIGMOD Record* (Morgan Kaufmann) 31, n° 1 (2016).

Yiou, Pascal. «Exploitation des bases de données pour évaluer les impacts du changement climatique.» Dans *Le développement durable à découvert.*, de Agathe, et al., ed. Euzen, 282-283. Paris: Éditions, CNRS, 2013.

Liste des tableaux et figures

Figure 1: Statistique de l'évolution du volume de données créés.....	9
Figure 2: Hadoop's System	11
Figure 3 : Exemple d'utilisation du modèle Map Reduce, afin de compter le nombre d'occurrence d'un mot dans un texte donné.....	12
Figure 4: Représentation de l'effet de serre	14
Figure 5: Centrales qui sont refroidies par les cours d'eau, répercussion sur la production d'électricité.	15
Figure 6: Data center de Google en Finlande.....	17
Figure 7: Cadre du big data dans les études sur le changement climatique. (Hossein Hassani 2019).....	21
Figure 8: Processus de la télédétection passive.....	
Figure 9: Présentation du Laboratoire virtuel BCCVL (Willow Hallgren 2016)	25
Figure 10: Surveillance de la qualité de l'air par les citoyens du canton de Liberty (Jennifer Gabrys 2016)	28
Figure 11: Structure des techniques clés du Big data. (Hossein Hassani 2019).....	30
Figure 12: Liste des sources de données climatiques (James H. Faghmous 2014).....	33
Figure 13: Cadre proposé pour la détection du changement climatique. (Gunasekaran Manogaran 2018).....	36
Figure 14: Architecture du pilote (Albani 2016).....	41
Figure 15: Ratio du « réchauffement de la planète » par rapport au « changement climatique » par cadres dans quatre pays. (S. Mo Jang 2015).....	44