CONSERVATOIRE NATIONAL DES ARTS ET METIERS

CENTRE REGIONAL DE RHONE-ALPES

Mémoire présenté en vue d'obtenir
UE « Information et communication pour ingénieur »
Spécialité : INFORMATIQUE

par

Galsungen

Big Data en médecine, en smart cities… Principes, utilités, exemples et solutions

Soutenu le 10 juin 2014

JURY

PRESIDENT	**[Civilité Prénom NOM *Fonction*]**
MEMBRES	**[Civilité Prénom NOM *Fonction*]** **[Civilité Prénom NOM *Fonction*]**

Abréviations

API : « Application Programming Interface » ou interface de programmation. Ensemble normalisé de classes, méthodes, fonctions servant de façades afin qu'un logiciel offre des services à d'autres logiciels.

GPGPU : General-purpose **p**rocessing on **g**raphics **p**rocessing **u**nits. Calcul générique sur un processeur graphique, afin de bénéficier de la capacité de traitement massivement parallèle.

HDFS : Hadoop Distributed File System

PSI : Public Sector Information ou Informations du Secteur public

RDF : Resource Description Framework. Modèle de graphe décrivant de façon formelle les ressources Web et leurs métadonnées, afin de permettre leur traitement automatique.

SAN : Storage Area Network

SIG : Système d'**i**nformation **g**éographique. Système d'information qui permet de créer, organiser et présenter des données alphanumériques spatialement référencées (géoréférencées). Les représentations sont souvent en deux dimensions (cartes) mais des représentations en 3D sont possibles.

SPARQL : SPARQL **P**rotocol **a**nd **R**DF **Q**uery **L**anguage. Langage de requête et protocole qui permet de rechercher, ajouter, modifier ou supprimer des données RDF.

Glossaire

Cloud : cloud computing, nuage. Ensemble de processus consistant à utiliser la puissance de calcul et/ou de stockage de ressources informatiques à travers un réseau.

Domotique : C'est l'ensemble des techniques de l'électronique, de physique du bâtiment, d'automatisme, de l'informatique et des télécommunications utilisées dans les bâtiments, plus ou moins « interopérables ». Elles permettent de centraliser le contrôle des différents systèmes et sous-systèmes (chauffage, électricité, alarme, communications, ouvertures…). Le but est d'apporter des solutions techniques pour répondre aux besoins de confort, de sécurité et de communication dans un bâtiment.

Épidémiologie : Étude des facteurs influant sur la santé et les maladies de populations. Discipline se rapportant à la répartition, à la fréquence et à la gravité des états pathologiques.

Écoépidémiologie : L'épidémiologie environnementale est une discipline émergente, transversale aux domaines de l'écologie et de la médecine, qui cherche à comprendre comment les modifications environnementales des activités humaines influent sur la santé humaine ou animale et s'intéresse donc aux relations écologiques entre les facteurs de pathogénicité, les populations cibles et l'environnement.

Informatique ubiquitaire : ubiquitous computing. Ère de l'informatique où nous avons à disposition des appareils (ordiphones, assistants personnels, capteurs…) qui font partie de notre quotidien et qui peuvent enregistrer, échanger ou permettre la consultation de données de façon permanente, quelle que soit la position géographique.

Scalable / **scalabilité :** Terme qui désigne la capacité d'un produit à s'adapter à un changement d'ordre de grandeur de la demande (montée en charge). Il n'y a pas de réel équivalent admis en français aussi utilise-t-on le terme anglais ou sa francisation populaire, ou encore le terme échelonnabilité. La scalabilité fait donc référence à la capacité d'un système à accroitre sa capacité de calcul sous une charge accrue quand des ressources sont ajoutées.

Smartphone : ordiphone ou téléphone intelligent. Téléphone mobile évolué disposant des fonctions d'un assistant numérique personnel, d'un appareil photo numérique et d'un ordinateur portable.

Sommaire

Liste des figures

Introduction

Les données de masse, plus communément nommées Big Data, résultent de l'évolution croissante des données. Ces dernières années, elles sont devenues plus communes par une démocratisation accrue. Et pourtant il est souvent fait des amalgames entre les grosses bases de données, la BI (Business Intelligence) et le Big Data.

Pour beaucoup, il s'agit des données générées par les réseaux sociaux ou celles enregistrées par les moteurs de recherches des grands du Web et de leurs produits associés. Quelques-uns ont conscience d'une dimension supplémentaire avec l'enregistrement des données d'utilisation des ordiphones (smartphones) et tablettes. Peu sont ceux qui appréhendent ce qu'elles représentent réellement.

En effet, au-delà de ce qui est connu, il s'agit d'un vaste système passant par de nombreux aspects : appareils informatiques et capteurs divers (informatique ubiquitaire), leurs utilisations dans les villes ou au sein de nos demeures et entreprises (domotique), la récolte des informations par tous les acteurs de notre quotidien (électricité, gaz, téléphonie, gouvernement, communes, transports…), l'évolution annoncée vers un Internet des objets… Les sources sont vastes et variées. La presse numérique de ce printemps mentionne que nous aurions généré plus de données cette dernière année que dans l'histoire de l'humanité. Eric Schmidt, le PDG de Google, dit lui-même que tous les deux jours nous générons autant de données que depuis 2003.

Mais que sont donc les Big Data ? Et quel est notre intérêt à générer de tels volumes de données ? C'est ce que nous allons développer ici.

Dans une première partie, nous définirons plus précisément ce que sont les données de masse (Big Data) et les technologies utilisées pour traiter, analyser ces volumes de données.

Ces explications seront ensuite illustrées à travers des exemples contemporains, plus particulièrement sur leurs usages en médecine, mais aussi au niveau des villes intelligentes (smartcities).

Enfin, afin d'ouvrir le débat nous parlerons de la législation sur l'utilisation de ces données, mais aussi de certains problèmes déontologiques et sociétaux qu'elles génèrent par rapport au respect de la vie privée ou au droit à l'oubli, avant de conclure.

1. Principes et technologies

1.1. Définition : Que sont les Big Data ?

Le Big Data ou grosses données est une notion apparue avec la multiplication des données produites et l'évolution des moyens de stockage. Les nouveaux médias et nouvelles technologies des vingt dernières années, ont permis la naissance d'une nouvelle richesse, les données. Nous pourrions les qualifier d'or noir numérique, nécessitant un stockage, traitement et raffinement afin d'être utilisées, valorisées. Un des premiers exemples est la société Google qui a dû faire face à un volume de données toujours croissant au travers de l'indexation de la toile, mais aussi par le stockage et l'étude des comportements des internautes, afin de proposer de nouveaux produits ou usages par segments.

La multiplication des technologies et surtout la démocratisation de certaines comme l'usage des téléphones portables (notamment les smartphones), des réseaux sociaux (Facebook, Twitter, LinkedIn…), des caméras de vidéo surveillance ou tous les capteurs et autres objets interconnectés, qui se multiplient, ont participé à l'expansion rapide des données et à la modification des usages.

Le volume est tel que les bases de données traditionnelles sont inadaptées, tant par leur capacité que par leur modèle relationnel. De plus, la nature des données, de plus en plus variée, et leur vélocité (rapidité de création, d'acquisition…), demandent aussi des traitements différents, plus dynamiques. Les données de masse se composent aussi bien d'enregistrements dans des bases NoSQL, que de documents vidéo, audio ou sous divers formats.

Si l'évolution technologique a contribué à leur émergence, elles ont provoqué en retour des évolutions et des retours sur investissement. En effet leur exploitation a nécessité de nouveaux besoins en calculs parallèles, en scalabilité et en stockage distribué.

Lors d'un rapport de recherche de l'institut Gartner par [Laney-2001], il a été décrit une règle dite des 3V pour qualifier ce phénomène.

- Volume
- Variété
- Vélocité

Deux autres V reviennent aussi souvent, dans les documentations :

- Véracité ou validité
- Valeur

Quelques documentations tentent de remplacer la validité par la visibilité, non pas dans un sens de simplification des données, mais de compréhension.

Le Big Data est une des voies de développement technologique les plus prometteuses dans la décennie à venir. Les gouvernements et administrations publiques s'en emparent aussi, que ce soit pour répondre à des problématiques de diffusions de l'information, de transparence, ou de valorisation des données publiques par la création de nouveaux produits. Plusieurs projets dits d'Open data

ont d'ailleurs vu le jour. Nous pouvons citer la mission interministérielle Etalab en France, ou le site de « Smart Data » pour la ville de Lyon, par exemple.

1.2. Quelles technologies ?

Mettre en place le Big Data, ce n'est pas simplement choisir une technologie, mais plutôt lancer une démarche, une réflexion sur ce que l'on veut obtenir des données, sur le résultat souhaité et sur les moyens pour y parvenir. Il n'y aura donc pas un outil spécialisé, mais plusieurs et ce, même si certains semblent proposer des boites à outils assez fournies. Citons comme exemple le plus probant, Hadoop de la fondation Apache [Apache Hadoop - 2014]. Cette solution est actuellement la plus utilisée lors de la mise en place d'architecture en Big Data. Ci-dessous une illustration des principales briques applicatives composant cette « boite à outils ».

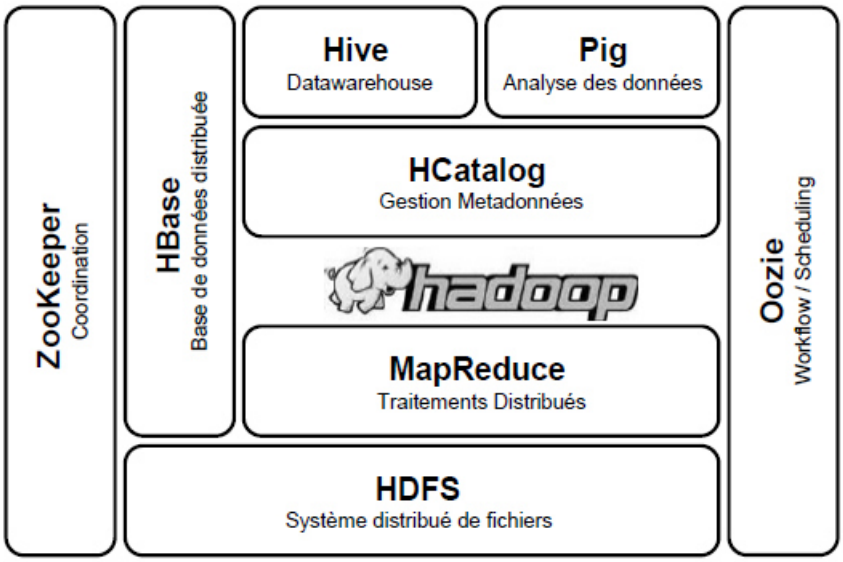

Figure 1 : Principaux composants d'Apache Hadoop [Brasseur-2012]

Bien sûr, bien d'autres briques, disponibles sur le portail Internet, peuvent être intégrées suivant les besoins. L'ensemble Hadoop reste la solution de référence en matière d'outillage pour les données de masse. Beaucoup de solutions de gros éditeurs s'appuient d'ailleurs sur certaines briques. Il en va ainsi de Azure, la solution « Cloud » de Microsoft qui, bien qu'il propose des bases de données en SQL Serveur, s'appuie sur un système de fichier HDFS.

Dans l'ensemble, nous pouvons ranger les logiciels en trois catégories [Chalmers et al. - 2013] :

- Stockage
- Traitement & calcul
- Analyse

1.2.1. Stockage

Le stockage, longtemps limité par la capacité des machines, s'est amélioré avec la forte augmentation des capacités des disques durs, mais aussi avec la démocratisation des baies SAN, des disques SSD (solid-state drive) et surtout du Cloud. C'est ce dernier aspect qui est le plus développé aujourd'hui avec une organisation des serveurs en grappes ou réseaux de nœuds. Il devient ainsi possible

d'ajouter facilement de nouveaux nœuds (ou nodes) à l'ensemble, suivant les besoins. C'est ce qui permet aussi, par la répartition ou copie des données d'offrir une tolérance de panne, de perte, par rapport à ces données. Au niveau de Hadoop, HDFS permet de partager l'espace nécessaire. GoogleFS ou Amazon S3 sont dérivés de ce dernier. Ce type de stockage est nécessaire quand on ne sait pas quelle taille pourra atteindre l'entrepôt de données ou que la nature des données est très variée.

De même, les bases de données NoSQL (Not Only SQL) ont connu un succès croissant ces dernières années. On peut distinguer quatre principaux types de base NoSQL :

- Paradigme clé/valeur
  - Exemples : Redis, Riak, Voldemort
- Bases documentaires
  - Exemples : MongoDB, CouchDB, Terrastore
- Bases orientées colonnes
  - Exemples : Cassandra, Amazon SimpleDB, Google BigTable, HBase
- Paradigme graphe
  - Exemples : Neo4j, OrientDB

Ces bases sont notamment utilisées par des géants du Web comme Google, Amazon, Facebook ou Ebay car elles offrent une plus grande simplicité par rapport aux bases relationnelles, mais aussi une meilleure montée en charge (scalabilité) par la multiplication du nombre de serveurs. Certaines ont été directement conçues pour être distribuées et ainsi éviter un point de défaillance unique (Single Point Of Failure ou SPOF). La solution Cassandra, née dans les locaux de Facebook et maintenant sous l'égide de la fondation Apache, est ainsi organisée en anneau avec plusieurs nœuds, l'ajout d'un nœud étant très simple.

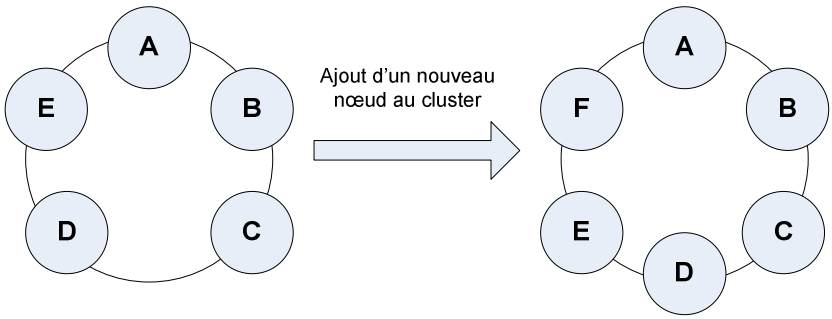

Figure 2 : Schéma présentant l'ajout d'un nœud ou node au sein d'un cluster [Foucret-2011]

La figure ci-dessus illustre un cluster de cinq nœuds auquel s'en rajoute un, simplement. L'ajout n'est pas fait suivant un ordre et le nouveau nœud peut très bien être placé entre A et B, ou D et E. Cela n'a pas d'importance. Il est donc possible d'augmenter ou de réduire la taille du cluster dynamiquement en fonction des besoins.

En complément, la figure ci-après présente le principe de réplication entre clusters, qu'ils soient hébergés au sein du même centre de données (datacenter) ou sur un site distant. Les quatre grappes (clusters) sont réparties entre deux centres d'hébergement. Une réplication est faite au sein de chaque centre, ainsi qu'entre les

centres. C'est ce qui va permettre aussi bien une tolérance de panne que l'ouverture des possibilités pour faire de la haute disponibilité.

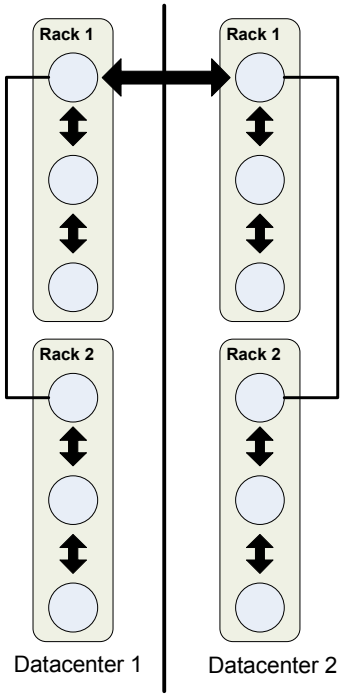

Figure 3 : Réplication de serveurs (nœuds) entre clusters et datacenters [Foucret-2011]

Un cluster n'est pas obligatoirement un ensemble physique. En effet, nous pouvons avoir des serveurs répartis sur différents centres d'hébergement distants. La grappe sera donc construite à travers un réseau plus vaste et les débits seront à prendre en compte dans la mise en place de la solution. En effet, avec un volume de données trop important, et surtout une vélocité forte, le réseau pourrait devenir le point de blocage, le frein, de l'architecture mise en place.

1.2.2. Traitement & Calcul

Les clusters (grappes de serveurs), le Cloud et les cartes graphiques sont les évolutions les plus récentes dans ce domaine. Cela a permis une augmentation des capacités de calcul, aussi bien par l'évolution des processeurs que par l'utilisation des multiprocessus et des traitements parallèles. Il est maintenant possible de louer des ressources de calcul dans le nuage. Par exemple, Amazon et Microsoft proposent des offres dans ce domaine.

De même, les évolutions technologiques nous ont offert des moteurs de calculs pouvant fonctionner seuls ou en grappes (clusters) comme à travers Hadoop ou Twitter Storm. Mais nous avons aussi assisté à la migration des calculs dans les processeurs graphiques, comme avec Nvidia CUDA et la programmation GPGPU. Dans ce cadre, au lieu de répartir les calculs entre les quatre cœurs du processeur machine, nous allons pouvoir les soumettre au processeur de la carte graphique.

Par exemple, une Nvidia GTX 780 contient 2304 cœurs [GEFORCE - 2014]. CUDA propose donc une API pour servir de pont entre l'hôte et le périphérique afin de pouvoir y passer les données et fonctions à exécuter sur le processeur graphique et lire les résultats en retour. Le plus gros défaut de cette technologie est qu'elle nécessite l'écriture complète d'un programme pour réaliser le processus et ce, dans

les langages C/C++, qui sont les seuls acceptés par CUDA. Cela implique donc un coût annexe dans leur utilisation. De plus, les processeurs graphiques sont plus adaptés aux calculs numériques notamment à virgule flottante. Il vaut donc mieux les conserver pour ce type d'opérations. Si cette solution est propriétaire, des libres ou Open sources sont disponibles comme *l'architecture* (framework) OpenCL, compatible avec les processeurs NVIDIA et AMD.

D'autres solutions intervenant dans les traitements et calculs existent comme « Bulk Synchronous Parallel Processing » (*modèle de construction* (bridgin model) pour la création d'algorithmes parallèles et fournissant un lien conceptuel pour la communication et synchronisation entre matériel et logiciel), « GraphLab » (une architecture logicielle de calcul distribué haute performance basée sur des graphes), « Disk-Based Graph Processing » (applications de traitement des graphes et API correspondantes, fournissant une capacité de traitements importante sur une machine unique en utilisant ses disques durs). De même, nous trouverons de nombreuses solutions développées pour un usage spécifique, afin de s'adapter de plus près au besoin.

1.2.3. Analyse

Il s'agit surtout du travail réalisé par les équipes et notamment par les statisticiens ou analystes des données (*data scientist, data analyst…*). Il s'agit donc d'un travail de réflexion pour exploiter au mieux l'ensemble des données et en tirer les informations ou conclusions nécessaires. Les outils sont donc moins nombreux et se présentent sous la forme de bibliothèques de fonctions ou d'apprentissage automatique, pour accompagner dans son analyse le scientifique des données.

Citons deux solutions, à savoir Mahout, bibliothèque d'apprentissage automatique (machine learning) et de forage des données (data mining) incluse dans Hadoop, et MLPACK. Cette deuxième, est une librairie d'apprentissage automatique et scalable en C++. Ces librairies offrent des algorithmes et méthodes exploitables au travers d'API.

1.2.4. Bilan technique

Le Big Data n'est pas lié à une solution technique, mais plus à un ensemble d'outils. C'est donc davantage une réflexion générale sur le besoin et son exploitation. Sont choisis ensuite les outils nécessaires à la réalisation du besoin. Le choix de la solution technique dépendra donc énormément des données étudiées, du besoin ou problème posé, de l'organisation et de la nature des données…

Ces outils sont bien souvent encore jeunes dans leur développement. En revanche, ils présentent une évolution rapide. Ainsi a-t-on vu Hadoop passer d'une interface en ligne de commande, à une interface au sein d'un navigateur, plus proche du besoin utilisateur. La tendance est vers une offre plus riche, voire aussi aboutie que celle présente pour la BI (Business Intelligence).

2. Utilités et exemples

L'analyse des données de masse touche tous les domaines ou presque, entre autres le marketing pour l'analyse ou la segmentation de la clientèle, la recherche pour le traitement, la vérification et la compilation de données, l'information et la recherche

d'information, le pilotage d'entreprise ou des finances, la prévention du crime… Un rapide tour d'horizon sur les usages du « Big Data », permettra ensuite de s'attarder sur des utilisations en médecine et dans les villes intelligentes (smartcities).

2.1. Usages du Big Data

Quand on parle de Big Data, Google, Facebook, Twitter, Apple ou Microsoft sont vus comme les précurseurs. Il est vrai que ce sont les sociétés qui ont le plus contribué à les faire connaître et que les volumes qu'elles gèrent et génèrent sont conséquents.

Au niveau d'Internet, les données de masses pourraient être considérées comme nées avec les cookies qui enregistraient et transmettaient des données au site les ayant générés. Cela s'est multiplié par la suite, et la traduction la plus fréquente de leur usage se présente sous la forme de publicités ciblées ou d'un filtrage des recherches suivant nos intérêts. Cela peut-être constaté en utilisant Google. La même recherche avec un compte ou en anonyme (navigation privée), présentera une différence entre les résultats. De même, en naviguant sur Google ou Facebook, nous pouvons voir des publicités orientées en fonction de notre historique de navigation ou d'achat. Amazon le traduit bien sur son site marchand avec des suggestions de produits pouvant plaire aux clients en fonction de leurs achats, mais aussi de leur navigation sur le site.

Même les gouvernements s'y sont intéressés. L'exemple le plus connu, récemment, est le projet PRISM de la NSA, qui a parcouru, et analysé une masse de données immense, pour la plupart issue des géants du Web. La NSA a ainsi pu dresser des profils de la population, la segmenter, mais aussi la surveiller de manière indirecte.

2.1.1. Cartographies et SIG

Les systèmes d'information géographique ne sont pas nouveaux. Depuis bien longtemps, nous les trouvons en agriculture, en foresterie, en écologie, démographie, géologie… Il s'agit de réaliser des cartes avec des données alphanumériques géoréférencées. L'évolution de l'informatique et la multiplication des données, permettent de créer de plus en plus de cartes et de les préciser. Des organismes proposent même des entrepôts de données accessibles par tous, via de simples flux. Le portail Smart Data de la ville de Lyon en est un exemple.

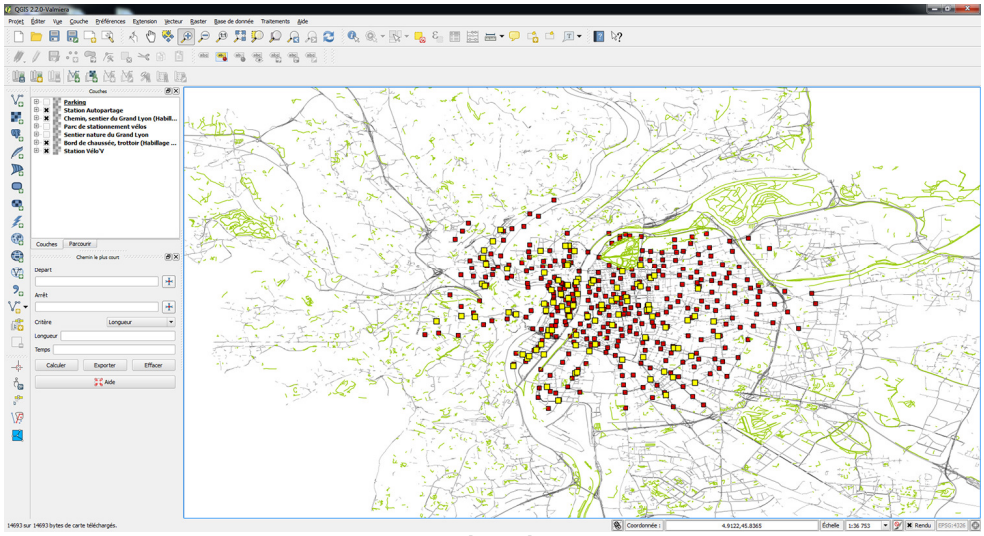

Figure 4 : Exemple de carte SIG réalisée avec QGIS et Lyon Smart Data

La figure précédente nous présente un exemple de carte réalisé avec un logiciel de SIG connecté à la plateforme de Lyon Smart Data. Y sont représentées les données suivantes : stations d'autopartage (carrés jaunes), stations Velo'v (carrés rouges), Chemin et sentiers du Grand Lyon (traits verts) sur une trame des bords de chaussées et trottoirs (traits gris). Il s'agit de quatre flux se superposant.

2.1.2. Prévisions politiques / sportives

Un autre domaine bénéficiant des données de masse est celui des prévisions, qu'elles soient dans le domaine politique ou sportif. Ici, on travaille sur les données avec des algorithmes prédictifs afin de prédire un résultat en fonction d'éléments présents, passés, environnementaux…

Un des exemples les plus connus est celui de Nate Silver, un statisticien américain, qui a réussi à prédire, avec une très grande précision, le résultat de l'élection présidentielle américaine de 2008 dans 49 des 50 états, et à le prédire pour les 50 états en 2012. Ce sont d'ailleurs les prédictions de Nate Silver en 2008 qui ont fait que Barack Obama, lors de sa campagne de 2012, ait mis en place une cellule spécialisée dans les Big Data et la segmentation de son électorat. Depuis Nate Silver a changé de domaine pour un plus lucratif qui est celui des prévisions sportives, et notamment, des ligues majeures de baseball et de football américain. Il s'agit d'un secteur ayant eu une forte expansion ces dernières années, plusieurs entreprises s'étant créées, chacune avec son propre algorithme prédictif.

On essaye, à l'aide d'algorithmes, de prédire les résultats des évènements sportifs en analysant les évènements passés. Dans ce cas, on analyse des entrepôts de données contenant aussi bien des informations sur la composition actuelle des équipes, mais aussi les résultats et informations des joueurs, ainsi que les données d'équipes ou de rencontres similaires. L'algorithme est censé donner un résultat assez précis. Nous sommes encore loin de pouvoir prédire facilement tous les résultats, mais certaines entreprises arrivent à faire des prédictions avec une précision de 75 % sur une année.

2.1.3. Marketing & publicité

Ainsi que mentionné avec les géants du web et leurs régies publicitaires associées, les données de masse ont un très fort usage en marketing et publicité, par l'analyse de la clientèle et sa segmentation. Si la première traduction est l'affichage de réclames, ou de suggestions ciblées, quand nous naviguons sur la toile, il y a bien sûr aussi des usages dans la vie courante et dans les orientations de la publicité urbaine ou télévisuelle. Il ne s'agit pas d'un phénomène nouveau, car depuis longtemps déjà des fichiers clients se vendaient, s'échangeaient afin d'élargir le dossier client en vue de démarchages ou de segmentation. C'est juste passé à une échelle supérieure avec des fichiers de plus en plus gros et contenant de plus en plus de données. Bien souvent, les données de masses ne permettent pas une utilisation directe, et une analyse est nécessaire. Dans ce cadre, ce sont plus des algorithmes descriptifs qui seront utilisés pour faire parler ces données.

Une nouvelle manière d'adresser la publicité joue sur les éléments déclencheurs. Par exemple, la société KXEN basée à San Francisco analyse les données géolocalisées des dépenses par cartes bancaires, afin de comprendre les comportements de clients. En couplant ces informations avec la position d'un client,

donnée par son smartphone, ils peuvent alors transmettre au téléphone des offres spéciales quand ce dernier passe à proximité de l'enseigne [Spécial Investigation - 2014]. Cela donne une publicité plus ciblée, perçue comme moins intrusive par les clients, et favorisant les achats immédiats pour profiter des offres. Les analyses tendent aussi à prendre en compte les humeurs des clients, via les réseaux sociaux, afin de trouver le bon moment pour leur faire des suggestions. Il ne s'agit plus seulement de chercher à analyser les comportements, mais à les influencer en fonction des humeurs.

2.1.4. Sciences

Les sciences ont toujours représenté des domaines pouvant générer de très gros volumes de données. Et les évolutions technologiques continuelles, comme les détecteurs de plus en plus sensibles, l'augmentation des capacités de stockage et du pouvoir de traitement des données, n'ont fait qu'accentuer ce phénomène. Certains domaines sont explicites sur ce point comme l'astronomie ou la cosmologie. Citons l'initiative « Square Kilometer Array » située en Australie et en Afrique du Sud. Ce réseau de milliers de télescopes va produire 700 téraoctets d'images chaque jour [IIP Digital - 2013]. Et ce n'est qu'un exemple, si l'on se base sur le nombre de télescopes et satellites actuels ou à venir. La météorologie, depuis des années, analyse de forts volumes de données pour donner des prévisions météorologiques grâce à l'analyse prédictive, et pour comprendre le système climatique mondial.

Parallèlement, nous assistons à une ouverture de vastes domaines de données dans le cadre de la recherche mondiale. Par exemple, même si tous les génomes séquencés actuellement ne sont pas en libre accès, un échantillon important a été mis à disposition pour aider les chercheurs et les universités. C'est le mouvement « open science data » dérivé lui-même du mouvement « Open data ».

2.1.5. Données ouvertes (Open data)

Une donnée ouverte ou open data, est une donnée numérique, issue du domaine public ou privé. Produite par une collectivité ou un service public, elle est diffusée de façon structurée selon une méthodologie et une licence (ouverte) garantissant son libre accès et sa réutilisation par tous, sans restriction (technique, financière, juridique…).

Ce concept est arrivé par le milieu scientifique en 1957-58. Il aura fallu attendre que les technologies permettent sa réelle exploitation et démocratisation. Nous pouvons donc assister ces dernières années à la multiplication des plateformes. Les gouvernements s'y sont intéressés, car cela permet de mettre à disposition du public des informations sur les institutions et ainsi répondre à certaines nécessités de transparence. Le projet data.gov lancé aux USA en 2009, et son homologue français, data.gouv.fr lancé en France en 2011 par la mission interministérielle Etalab en sont des exemples.

Cette ouverture des données publiques n'est pas nouvelle, car elle se trouve dans l'article XV des droits de l'homme. Aux USA une loi de 1966, dite « Freedom of Information Act » promouvait d'ailleurs cette ouverture. En France nous ne sommes pas en reste, car en extension de la Déclaration universelle des droits de l'homme, il y eut plusieurs textes de loi dans ce sens. Citons la directive PSI, qui fût traduite par une ordonnance, puis par un décret du 30.12.2005, extension de la loi CADA

(Commission d'Accès aux Données Administratives) de 1978. Au niveau européen nous pouvons citer les directives 2003/98/CE du 17.11.2003 et INSPIRE de 2008 qui encadrent ce phénomène.

Plusieurs moyens sont disponibles pour évaluer l'ouverture des données. La Sunlight Fondation a établi une liste de 10 critères [Sunlight Foundation-2010] afin de caractériser ces dernières :

- Complète
- Primaire
- Opportune
- Accessible
- Exploitable
- Non discriminative
- Non-propriétaire
- Libre de droits
- Permanente
- Gratuite

[Tim Berners-Lee-2010] a aussi publié une échelle de qualité des données ouvertes avec une notation d'une à cinq étoiles pour réaliser cette évaluation.

Le phénomène d'open data n'est pas qu'un effet de mode. La mission Etalab a, par exemple, élargi l'offre française au printemps 2014, avec le lancement d'une plateforme dédiée à l'enseignement supérieur et à la recherche et proposant déjà 23 jeux de données [Lemaire-2014]. D'autres organismes lancent aussi des portails de données ouvertes comme Wikipédia avec le projet Wikidata.

Cette ouverture donne naissance à une économie des applications et à une explosion des communautés de développeurs. En revanche, il faut s'assurer de conserver l'interopérabilité des données, notamment avec le format RDF et le langage SPARQL, par exemple. L'homogénéisation du format de description des ressources ou données, et du langage permettant de consulter ces descriptions simplifie les interactions entre différents acteurs en offrant une base commune de travail et d'échange. Également, s'assurer que les licences sous lesquelles sont publiées ces données sont bien compatibles avec le but final. Cela pose souvent un problème en France et donc une nécessité d'anonymisation des données pour rester dans le cadre légal.

2.2. Le Big Data et la médecine

Comme abordé précédemment, les sciences, et notamment la médecine, ont fortement bénéficié des apports du Big Data, et pas uniquement au travers du mouvement « open science data ». L'enregistrement électronique des dossiers avec les actes médicaux et observations, mais aussi l'augmentation du nombre d'appareils connectés liés à la médecine (bracelets, capteurs, senseurs, imagerie, séquençage génétique, analyses biologiques…) génèrent une masse de données à traiter. Si dans certains cas cela peut-être fait directement, il y a deux écueils fréquents. Le premier est l'adaptation nécessaire des médecins et personnels hospitaliers à ces technologies, mais aussi leur acceptation que le patient ait un rôle de plus en plus actif dans l'étude de son propre traitement. Le second est bien sûr la diffusion de données médicales et donc personnelles par les patients, sans qu'ils n'y voient ou n'y subissent une atteinte à leur vie privée [Ranck-2012].

2.2.1. Epidémiologie & Ecoépidémiologie

La multiplication des données dans le milieu hospitalier est devenue un fait et permet une analyse de masse des facteurs influant sur la santé et les maladies. Cela peut permettre, par une segmentation des personnes, l'identification des publics plus sensibles à certains maux et ainsi de les traiter préventivement. Une meilleure répartition des ressources comme les vaccins serait envisageable. Ainsi celui de la grippe pourrait n'être proposé qu'à des personnes présentant une réelle sensibilité, un réel risque.

Le couplage des données médicales avec d'autres, environnementales, météorologiques, géographiques, anthropologiques… offre aussi de nouveaux degrés de compréhension pour certaines maladies infectieuses ou parasitaires : pourquoi certains maux apparaissent et, à l'avenir, comment les prévenir. L'augmentation des pluies, ou d'une saison humide, peut générer une augmentation des populations d'insectes (comme les moustiques) et ainsi augmenter les risques de propagations de certaines maladies. Ce modèle peut aussi être adapté lorsque des catastrophes naturelles se produisent.

Au-delà de la compréhension, il est possible d'imager ces phénomènes à travers des représentations géographiques. Citons le [Réseau Sentinelles - 2013], qui diffuse des cartes et des informations hebdomadaires sur l'état des épidémies en France. Dans le même domaine, Celipharm traite chaque jour plus d'un million de données brutes provenant de tickets de pharmacies [Seibt-2014]. Ils ont ainsi créé le portail « Openhealth » affichant en temps réel des informations sur la santé des Français et des cartes en rapport (épidémies, allergies…). Certains adeptes du Big Data pharmaceutique n'hésitent pas à dire que ce réseau est plus à jour que l'officiel, étant donné qu'il travaille avec des données à J+1 contre un délai souvent d'une à deux semaines par le réseau habituel, le temps que les médecins transmettent leurs données. Cela offre donc une base sérieuse en épidémiologie pour la détection d'épidémies. Il est aussi possible de trouver d'autres débouchés à ces données comme l'analyse des prescriptions.

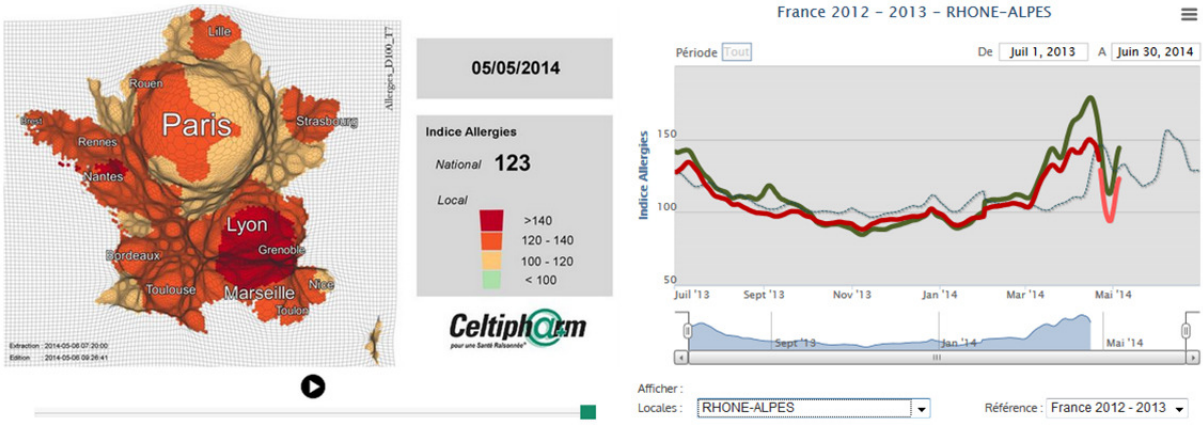

Figure 5 : Exemples de graphiques disponibles sur Openhealth.fr « incidence des manifestations allergiques ».

2.2.2. Séquençage génétique

Si le premier séquençage génétique a nécessité une dizaine d'années, il est maintenant possible d'analyser un génome en quelques semaines. Au-delà de l'amélioration des technologies pour analyser l'ADN, l'usage du Big Data a aussi

apporté sa participation à cette évolution. Tout d'abord, quand on séquence de l'ADN, le brin est divisé, puis analysé. De nombreuses informations sont recueillies, en général dans le désordre. Il est donc nécessaire de comparer la chaîne obtenue avec les génomes existants afin de pouvoir la positionner correctement. C'est ce qu'on appelle l'alignement des séquences.

Ensuite, un séquençage génétique comporte une partie de trous (de « gap »). Avec l'utilisation d'algorithmes prédictifs, il est alors possible de combler ces « trous ». Cela nécessite en revanche l'analyse des bases existantes et donc l'analyse d'un fort volume de données. Un humain compte normalement 46 chromosomes, soit 28 000 à 35 000 gènes, soit environ 3,1 milliards de paires. Cela nous donne une idée de la masse de données à analyser. Il est à noter que le but n'est pas uniquement de séquencer le génome humain, mais aussi celui de nombreuses espèces tant animales que végétales. De plus la tendance va maintenant au-delà du simple séquençage génétique, en tendant vers la connaissance et compréhension du phénotype [Meldrum et al.-2011], c'est-à-dire de toutes les variations génétiques caractérisant une même espèce. Dans ce cadre, les Big Data et la comparaison de schémas (patterns), servent aussi à identifier les variations, mutations naturelles ou non. C'est ce qui va permettre en cancérologie d'identifier les types de cancer et donc de proposer des traitements adaptés à chacun.

La bioinformatique a explosé ces vingt dernières années. Le premier logiciel ayant permis l'alignement des séquences dans les années 90 était BLAST. Il a depuis été dérivé en une version Cloud : CloudBLAST. Nous pourrions en citer d'autres, car ils sont des dizaines rien que pour ce domaine. Beaucoup de ces logiciels s'appuient sur des briques techniques que nous connaissons comme MapReduce ou Hadoop. Le logiciel CloudBurst qui permet de créer un index de k-mer (suites de bases) utilise MapReduce et MUMerGPU (parallélisations de calculs sur GPU comme avec Nvida CUDA) pour faire ses calculs. Ce ne sont ici que quelques exemples sur un écosystème large et varié. Autre exemple, l'université de Washington a diffusé SOAP (Short Oligonucleotide Analysis Package) ces dernières années [SOAP-2007] et ce, gratuitement pour les autres universités et laboratoires de recherches. Comme avec Apache Hadoop, nous avons là une boite à outils spécialisée dans l'analyse de la masse de données produites en séquençage génétique. Au-delà d'une explosion de l'informatique, il y a une transformation des métiers, car le chercheur doit maintenant avoir une culture informatique en plus de ses connaissances métiers. On parle d'ailleurs de bioinformatique.

Le mouvement open science data a aussi participé à cet essor. Au moins cinq bases de données ouvertes pour les génomes, dont le site 1000 Genomes, qui offre ses ressources aux chercheurs et universités. En revanche, la question de la confidentialité se pose ici. Sur le projet 1000 Genomes, le généticien Yaniv Erlich et son équipe ont réussi à identifier précisément cinq personnes masculines, en recoupant ces informations avec des données d'Internet, de généalogie… Bien sûr, ils n'ont pas publié leurs résultats pour tous et ont averti le projet pour qu'ils modifient les informations transmises sur les génomes. C'est donc une question d'éthique, de confidentialité des données et d'anonymisation. Il s'agit d'un des débats actuels sur ce sujet [Braly-2013].

2.3. Usage pour les villes intelligentes (smart cities)

La ville intelligente, du terme anglais smart city est un concept émergent décrivant un développement urbain moderne, communiquant et durable. Il pourrait être dérivé de la science-fiction et des descriptions faites par William Gibson ou Philip K Dick. Il s'agit d'une ville moderne et équipée d'infrastructures (eau, électricité, communications, gaz, transports, services d'urgences et de la santé, équipements publics, bâtiments…), communicantes et durables afin d'améliorer la vie des citoyens tout en étant efficient et en respectant l'environnement.

Même si nous n'en n'avons pas conscience, il s'agit d'un mouvement d'amélioration du quotidien qui est déjà en marche, et qui évolue tous les jours. Citons divers exemples, comme la mise en place de smartgrids (grilles électriques) afin d'optimiser la production électrique, mais aussi son approvisionnement et l'évolution de sa consommation, ou bien l'optimisation des transports en commun d'une ville, la favorisation de modes de transports doux ou électriques, l'amélioration des infrastructures réseau par le déploiement de fibres optiques… Les exemples ne manquent pas et tous ont souvent en commun un besoin initial de connaissances auquel peut répondre le Big Data. Dans un premier temps l'analyse des données collectées permet déjà de tirer quelques conclusions, mais en les recoupant avec d'autres comme celle de la téléphonie mobile, une compréhension bien plus vaste du problème posé est accessible. Dans un second temps, ce besoin étant croissant, l'objectif est de générer le plus de données possible en ajoutant des capteurs (comme sur les routes ou les canalisations d'eau, ou bien le déploiement de compteurs intelligents…).

2.3.1. Open data et smart data : exemple de Lyon

Comme nous avons pu le voir, entre autres avec la mission interministérielle Etalab et le lancement d'un portail d'Open Data public, de nombreuses villes ont profité de ce mouvement afin de libérer des données et de mieux répondre au rôle de transparence attendu de leurs populations. Le site opendata-map.org propose d'ailleurs une carte de ces communes, également listées sur la page Wikipedia dédiée aux données ouvertes en France. On y trouve Paris ou plutôt la région Ile-de-France, qui a également lancé un large projet, PRISME, et signé des partenariats avec l'INRIA, CITRIS et la ville de San Francisco afin de répondre à plusieurs objectifs [Inria-2014] :

- La démocratisation des données urbaines
- Les systèmes de transport intelligents
- La perception et l'analyse de la qualité de l'air et des émissions de dioxyde de carbone.

D'autres villes proposent des portails d'accès à leurs données comme Toulouse, Bordeaux ou Lyon à travers un portail nommé « Smart Data ».

Ce portail nommé « Smart Data Grand Lyon » permet d'accéder à 423 jeux de données actuellement. En constante évolution il se veut une brique dans une réflexion plus vaste de transformation de la capitale des Gaules en une ville intelligente et durable. Il ne s'agit bien sûr pas de l'ensemble des données du Grand Lyon, seulement de celles rendues publiques. D'autres jeux de données sont réservés à des acteurs publics ou privés dans le cadre de partenariats pour le développement d'applications, de services ou d'activités. Ceci se couple d'ailleurs

avec une plateforme d'expérimentation sur le quartier de la Part-Dieu permettant de tester ces réalisations innovantes en conditions réelles.

2.3.2. Gestion de l'énergie : les smartgrids

Ici aussi, nous nous trouvons devant une problématique mondiale. L'émergence des smartgrids, c'est-à-dire de réseaux électriques « intelligents », utilisant des technologies informatiques afin d'optimiser la production, distribution et consommation d'énergie et notamment d'électricité. Il ne s'agit pas non plus d'une nouveauté, car de longue date l'électricité a été organisée en réseaux comme en France où on a un découpage en 7 régions, mais d'une évolution de ce réseau pour prendre également en compte les nouveaux modes de production d'électricité, la détection de pannes et l'optimisation des consommations.

Ces points nécessitent de connaître les productions et usages afin de pouvoir mieux les répartir et aussi de pouvoir prévoir les montées en charge. C'est dans ce cadre qu'on a vu apparaître trois nouveaux types de compteurs :

- Compteur automatique : mesure, enregistre et transmet les données. À sens unique.
- Compteur communicant : Il transmet et reçoit des informations, permettant ainsi de communiquer avec le consommateur.
- Compteur intelligent : Evolution du compteur communicant afin d'aider le consommateur à piloter sa consommation.

Nous pouvons citer l'exemple de Linky en France qui est un compteur communicant. Il est actuellement déployé, en test, en Indre-et-Loire pour le territoire rural et sur Lyon pour le territoire urbain. En parallèle, une expérimentation est réalisée par la ville d'Issy-les-Moulineaux sur le quartier « Seine Ouest » afin de créer le premier « quartier intelligent ». Tous les logements ont donc été équipés d'équipements permettant de suivre les consommations par usage (chauffage, éclairage, eau…) ou par équipement à travers une interface. Les habitants ont, tout au long du processus, été intégrés à ces changements. En tout une dizaine de partenaires ont participé à ce projet IssyGrid qui s'appuie sur une infrastructure de données tournant dans le Cloud Azure de Microsoft [Brasseur-2013]. D'autres projets sont en cours, comme dans le quartier Confluences à Lyon où un ensemble de bâtiments intelligents est actuellement en construction avec le projet Hikari.

S'il faut parfois du temps pour mesurer les effets d'évolutions technologiques, les premiers résultats sur ce type de projet dans le monde montrent une baisse significative de la consommation et une forte amélioration du service. Bien qu'important, l'accompagnement des consommateurs pour la modification de leurs comportements n'est qu'une partie de la solution. Il faut aussi prendre en compte qu'il n'est pas question seulement de la construction de nouveaux bâtiments répondant aux normes, mais aussi de la transformation, de l'adaptation des bâtis existants.

2.3.3. Gestion & optimisation des transports

Au sein de toute ville, il y a une nécessité de déplacements pédestres et motorisés. Le Big Data peut aider à leur optimisation en apportant une connaissance accrue du trafic, des travaux, mais aussi des situations aux différents carrefours… Il sera donc possible d'étudier une réorganisation des transports publics comme l'ouverture de nouvelles lignes, ou bien le conseil d'itinéraires secondaires…

2.3.3.1. Exemple d'Abidjan

Abidjan en Côte d'Ivoire est une ville de 4,5 millions d'habitants. Son réseau de transport public est composé de 539 bus. Il est complété par des transports privés sous la forme de 5 000 minibus et 11 000 taxis. Comme la majorité des villes d'Afrique subsaharienne, ce réseau se détériore fortement. La ville a donc fait appel à une équipe dublinoise de la société IBM [Calabrese-2013] pour répondre à une problématique : comment mieux adapter le réseau de transport public aux besoins ?

Dans un premier temps, l'équipe a dû évaluer la mobilité des usagers. Les enquêtes sur le terrain étant trop limitées et coûteuses, elles ont été exclues. L'évolution de la téléphonie dans les pays d'Afrique se fait par une explosion des services mobiles. Par exemple, en Côté d'Ivoire, 70 % des habitants en sont équipés. Ils ont donc utilisé une base de données issue de la téléphonie mobile et fournie par l'opérateur Orange. La base de données contenait 2,5 milliards d'enregistrements (appels et SMS). Chaque enregistrement comprenait un identifiant anonymisé, l'heure à laquelle la communication a été passée ou reçue et l'identifiant de l'antenne-relais connectée au début. Après une première analyse de ces données, l'étude a porté sur 500 000 téléphones pour des appels passés en 2012, sur une période de cinq mois. Ils ont pu extrapoler les mouvements individuels des utilisateurs entre deux appels consécutifs. Cela a permis de donner une matrice origine/destination avec une représentation sur une carte présentée ci-après. En la comparant avec le plan de desserte actuel, ils ont pu mettre en exergue des zones avec de forts déplacements, mais non couvertes par le réseau public. Avec une deuxième analyse, ils ont pu quantifier et représenter la fréquentation des lignes existantes. Avec ces informations et l'aide d'un algorithme d'organisation, des propositions ont été faites afin d'ouvrir quatre nouvelles lignes, en rediriger d'autres…

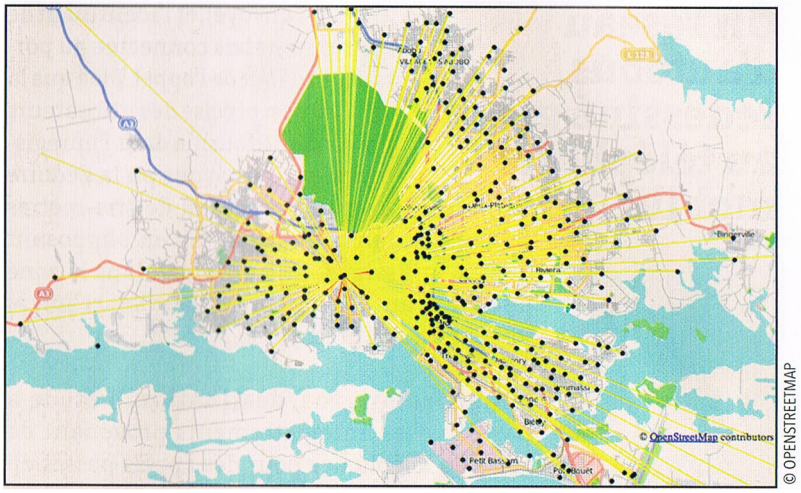

Figure 6 : Carte des flux moyens des déplacements par paires d'antennes-relais dans la ville entre 7 et 16 h, calculés avec les appels de 500 000 téléphones sur cinq mois.

Ce projet illustre bien comment les données de masse, ici issues de la téléphonie mobile, peuvent permettre de résoudre des problématiques d'optimisation des réseaux de transport. Cela nécessite bien sûr de pouvoir prendre du recul pour avoir une vue d'ensemble, mais aussi de ne pas se cantonner au domaine initial afin de trouver une solution.

2.3.3.2. Optimod & Onlymoov sur Lyon

Lyon, comme nombre de grandes métropoles, est confrontée à des soucis de congestion du trafic, ce qui freine la mobilité, mais aussi l'économie locale, tout en ayant un fort impact sur l'environnement. C'est dans ce cadre que plusieurs projets ont vu le jour. Le premier, s'appuyant, entre autres, sur les données de masse en extension du projet Smart Data et sur un système CRITER (Commande de Régulation et d'Information du Trafic et des Évènements Routiers) est Onlymoov. Installé dans l'hôtel de communauté du Grand Lyon, il bénéficie, en plus, du déploiement de capteurs (246 caméras mobiles, 1 255 carrefours connectés, 46 panneaux à affichage variable (PAV)), afin d'afficher en temps réel des informations aux usagers. Pour ce faire, en plus des PAV, il propose une interface Web et une application pour supports mobiles (ordiphones et tablettes).

En complément, un autre projet plus vaste, et englobant Onlymoov, est en cours de réalisation, il s'agit d'Optimod. Ici on souhaite prédire, réorganiser, réduire le trafic tout en offrant une mobilité et desserte maximale aux utilisateurs. Ce projet, qui se veut vraiment multimodal, a été découpé en neuf lots, chacun s'appuyant sur les précédents, mais surtout sur le traitement en temps réel des données de masse disponibles et remontées par de nouveaux capteurs, sources. Ce projet se veut plus large et non uniquement centré sur les usagers civils. Il a pour but de réduire la circulation et la pollution en ville tout en optimisant les tournées de livraisons, de fret, la navigation urbaine des poids lourds. Des partenariats ont été signés avec de nombreux acteurs comme IBM France dont la mission est ici de prévoir la circulation avec au moins une heure d'avance [Alternativ'City-2013]. Cette partie du projet, due courant 2014, avait déjà une précision de 90 % en 2013.

2.3.4. Lutte contre le crime (cas de Rio de Janeiro & de Los Angeles)

L'idée initiale est venue des soucis d'organisation, de coordination des secours suite à des catastrophes naturelles. Rio de Janeiro et Los Angeles sont deux exemples de villes ayant mis en place un centre de commandement des forces de l'ordre qui s'appuie sur l'exploitation de données de masses.

En analysant les données des crimes ou délits passés et en les couplant avec des algorithmes mathématiques, on peut distinguer des schémas qui vont aider à prédire les zones à surveiller particulièrement. On ne cherche pas à prédire les crimes, mais plutôt, en fonction d'un historique et d'événements, d'informations en temps réel, à prédire les zones à risque, c'est à dire où ils sont susceptibles de se produire. C'est ce qui a permis d'organiser les patrouilles des officiers de police. L'algorithme mathématique ici se rapproche de celui permettant de prédire les remous suite à une éclaboussure sur un plan d'eau. La multiplication des caméras et périphériques connectés aide à aller dans ce sens, par la profusion de données additionnelles qu'ils fournissent.

La ville de Los Angeles a constaté une diminution de 20 à 25 % des crimes suite à la mise en place de cette solution technique. Cela ne s'est pas fait si simplement, car il a fallu un temps d'adaptation assez long, et surtout convaincre les policiers de suivre ces conseils, de modifier leurs routines.

2.3.5. L'évolution des cités vers l'intelligence

Quand nous parlons de la création de villes intelligentes ou de smartcities, deux visions peuvent se confronter. La première est celle directement issue de la science-fiction avec des villes neuves construites en utilisant les nouvelles technologies. La deuxième, souvent moins démocratisée, mais pourtant sûrement la plus utilisée, est la transformation de nos villes en intégrant les technologies au sein de l'existant à chaque fois que cela est possible. « L'objectif est d'optimiser la ville sans forcément transformer le rapport à l'urbain » écrit Olivier Mongin dans un article de Libération [Féraud & Chetrit - 2013]. Cela va donc de l'installation de compteurs communicants ou intelligents, au fait de relier les quartiers par de la fibre optique, ou bien d'inclure dans les opérations de rénovation, de restauration, des systèmes domotiques ou énergétiques (isolation, panneaux solaires, pompes à chaleur…).

Les exemples ne manquent pas pour parler de ville intelligente. Il semble qu'une certaine « frénésie » se soit emparée de tous les pays sur cet aspect. À noter que même les sociologues ou anthropologues s'y intéressent maintenant. Ils permettent d'apporter une nouvelle dimension à ces études, comme celle de diriger l'évolution de ce concept vers un aspect plus anthropo-centré que techno-centré, c'est-à-dire en adaptant les technologies à l'homme et non en demandant à l'homme de s'adapter [Picon-2013].

2.3.5.1. Songdo : la nouvelle ville, intelligente

La ville de Songdo en Corée du Sud est le plus bel exemple d'une ville construite directement dans cette optique. Située à 65 kilomètres de Séoul, elle a pour ambition de devenir un nouveau pôle technologique et économique. Dès le départ, un soin a été apporté pour intégrer les capteurs, les éléments de domotiques et les moyens de communication directement dans les immeubles, aussi bien résidentiels que commerciaux. La fibre optique a donc été plébiscitée, mais aussi les moyens de visioconférences. Ainsi les salles de classe, les hôpitaux, les cabinets médicaux, mais aussi les logements des résidents, ont été équipés de caméras et d'écrans. Cela peut permettre à une personne malade de réaliser une consultation, directement avec un médecin ou un spécialiste, sans quitter son salon.

En parallèle, les rues, les croisements, les réseaux électriques et d'eau ont tous été équipés de capteurs afin d'être gérés en temps réel. Cette ville nouvelle a quand même du mal à attirer les entreprises. Même si elle se veut moderne, une grande partie des infrastructures a déjà dix ans, ce qui est important au niveau écart (*gap*) technologique, dans des domaines qui évoluent de plus en plus vite. Cela a aussi permis de mettre en exergue d'autres écueils comme le fait que des contrats d'exclusivité sur certaines technologies (partenariats datant des premières constructions), puissent devenir des freins à l'évolution rapide nécessaire pour suivre les nouvelles technologies [Marshall-2014].

2.3.5.2. Buenos Aires : la transformation d'une ville existante

À l'inverse de Songdo, beaucoup d'évolutions des cités actuelles vers les smart cities se font par la modernisation de bâtiments existants au cours de rénovations ou de travaux dédiés. C'est le choix qui a été fait par la ville de Buenos Aires. Cette transformation, commencée en 2007 suite à l'élection de Mauricio Macri, avait trois programmes parallèles : Gouvernement électronique (modernisation administrative), Ville ouverte (transparence des décisions) et Ville intelligente (amélioration de la

qualité de vie des habitants à l'aide des nouvelles technologies) [Buchet-2014]. Cela a commencé par l'informatisation des administrations et la formation de fonctionnaires. Par la suite ils ont mis en place un programme d'e-gouvernement. Ils ont donc rendu accessibles les démarches administratives, dossiers et informations publiques en ligne. Afin d'aider à la modernisation, tous les bâtiments publics proposent actuellement du Wifi en accès libre et gratuit. Le gouvernement a aussi doté les écoles et universités, et surtout les écoliers et étudiants, d'ordinateurs portables.

À l'aide d'entrepôts de données, des premières applications ont été mises en place comme :

- Applications mobiles diverses
- État de la circulation en temps réel
- Calculs d'itinéraires multiples
- Location de vélos et plan des pistes cyclables
- Agenda culturel
- Accès aux bibliothèques…

Afin de favoriser l'utilisation de ces applications, des bornes interactives ont été déployées dans toute la cité. Ces applications et leur utilisation génèrent à leur tour des données de masse qui permettent de les améliorer ou d'en développer de nouvelles. Récemment, ce programme s'est renforcé avec le déploiement de cartes à puces pour les transports en commun et aussi la mise en place de guichets décentralisés et automatiques pour délivrer des documents administratifs.

2.4. De nouveaux usages donnent de nouveaux métiers

Afin de traiter ces gros volumes de données, mais aussi de faire parler les données, que ce soit avec une méthode descriptive ou prédictive, un nouveau type de métier s'est fait jour. Il s'agit du « data scientist » (ou « data miner », « data analyst »), traduit en français par « analyste des données » (ou scientifique des données, statisticien, mathématicien). Les termes anglais restent les plus utilisés pour le moment, sans réel consensus pour leurs équivalents français.

Ce nouveau profil, souvent décorrélé de la DSI, nécessite des compétences avancées en mathématiques et statistiques, mais aussi en communication, des connaissances sur le côté métier et un aspect créatif. Il est possible de trouver dans quelques ouvrages des tableaux des compétences de base attendues. Il doit donc comprendre les problématiques du business, savoir collecter des données de multiples sources, qu'elles soient structurées ou non, optimiser la qualité des données, définir et utiliser des algorithmes, des modèles d'analyse, donner du sens aux données, mais aussi savoir communiquer. D'un point de vue généraliste, il devrait maitriser de nombreux champs en statistique (économétrie, probabilités, statistique inférentielle, analyse multidimensionnelle, séries temporelles, sondages…) et en économie (histoire économique, analyse d'entreprise, macro et micro-économie, économies du travail et industrielle, sociologie…).

De plus, le data scientist devra s'intéresser au métier du secteur donné. C'est-à-dire que pour analyser des données financières, il devra approfondir l'aspect finance. Bien sûr, des compétences en informatique sont obligatoires, afin de pouvoir accéder, gérer et traiter les données. L'aspect architecture technique, quant à lui,

reste au niveau de la DSI en général. Ce sont eux qui déploient et gèrent la partie technique et les nouvelles solutions. Le schéma suivant pourrait représenter les interactions entre les acteurs.

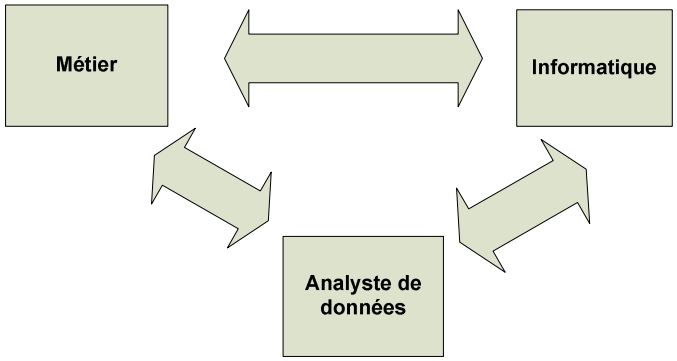

Figure 7 : Interaction entre compétences pour le Big Data

Comme nous pouvons le constater, l'analyste de données ne doit pas rester isolé afin de tirer le meilleur des données de masse, ces dernières possédant plusieurs axes, indispensables les uns aux autres.

3. Pour aller plus loin : Législation et vie privée

3.1. Confidentialité des données & vie privée

C'est un débat sur lequel il faudrait s'attarder. Déjà de nombreuses données ont été récupérées et analysées et notre ignorance ne peut justifier de ne pas nous y intéresser. C'est bien souvent sur un mode de consentement volontaire entendu que nous fonctionnons. Qui lit réellement les conditions d'utilisations des applications pour les ordiphones, ou pour les logiciels, les services en ligne… ? La plupart du temps, nous les acceptons, sans même les lire. Quand nous diffusons des informations sur Internet, nous n'autorisons pas pour autant qu'elles soient collectées par des tiers, sans accord préalable, afin d'être analysées.

La question n'est donc pas de cacher des informations ou sa vie privée, mais de pouvoir garder la main sur l'utilisation des données, et les intervenants à qui elles pourraient être communiquées. Nous pensons souvent que la législation de notre pays nous protège, mais ce n'est que rarement le cas. Ainsi nombre de données de masse sont étudiées dans d'autres pays, hébergées sur d'autres sols, sous d'autres contraintes légales. Les géants du Web comme Google, Amazon ou Microsoft ont leurs datacenter aux USA. Ils sont dépendants des lois américaines et tombent donc sous le « Patriot Act ». Cette loi autorise les services de sécurité américains à accéder aux données informatiques détenues par les particuliers et les entreprises, sans autorisation préalable et sans information des utilisateurs. La NSA et les autres services d'enquêtes ou de protections américaines ont ainsi accès à toutes leurs données, même si celles-ci concernent un ressortissant étranger. Plus proche de nous, les données des compteurs Linky d'ERDF, sont analysées aux USA par Teradata.

En France c'est bien la loi « Informatique et Libertés » du 6 janvier 1978 qui s'applique par rapport au fichage manuel ou informatique. La loi a bien sûr subi

quelques mises à jour en 2004, 2005 et 2007 sous la forme de compléments. Actuellement, elle pourrait se résumer aux points suivants :

- Le droit d'information (savoir si on est fiché)
- Le droit d'opposition (pouvoir s'opposer à notre fichage)
- Le droit d'accès (pouvoir accéder aux données nous concernant)
- Le droit de rectification (pouvoir modifier les données)

Ce n'est malheureusement pas toujours respecté, comme nous le verrons dans la partie dédiée au droit à l'oubli.

Un autre point concerne l'identification dite indirecte. Un groupe de chercheurs, en partie du MIT (Massachusetts Institute of Technology), a pu démontrer qu'il suffisait de 4 points, dans une base de données d'un opérateur de téléphonie, pour identifier 95 % des individus, contre 12 points avec des empreintes digitales [Montjoye et al. -2013]. Et il ne s'agit pas là d'un cas isolé comme nous avons pu le voir sur la partie dans le séquençage génétique. Ce risque est donc accru quand on recoupe les bases de données.

Dans certains cas, on parle d'une anonymisation des données, mais cela n'efface en rien le problème. En effet sur des bases de données médicales, il est possible avec les informations disponibles (commune, sexe, dates d'hospitalisation) de descendre rapidement dans la précision du ciblage. C'est d'ailleurs un point qui a été mis en exergue dans un rapport sur la gouvernance et l'utilisation des données de santé par Pierre-Louis Bras et André Loth. Ce rapport a remis en cause pas mal d'outils disponibles au niveau PMSI (Programme de Médicalisation des Systèmes d'Information).

Un autre argument, souvent avancé, est de baisser la granularité des données. Mais si pour une image il est possible de baisser la résolution, comment cela peut-il être possible avec des données ? En effet comme l'a montré l'étude « Unique in the Crowd », il suffira simplement d'ajouter un ou deux points, pour de nouveau identifier les personnes. Un autre aspect démontré par cette étude, est la nature des prédictions possibles, rien qu'avec les données issues des ordiphones. Une équipe du MIT a ainsi pu faire des prédictions précises sur des tests de personnalité comme le BFI (Best Five Inventory) et ce, en se basant sur les données des smartphones [Guillaud-2014].

Enfin, un dernier point, concernant la sécurisation de ces données. Même si nous donnons une autorisation pour les étudier statistiquement, il restera à prendre en compte la sécurisation sur le support informatique. Cela peut bien sûr passer par un chiffrement ou cryptage des espaces de stockages et des échanges entre machines, mais il restera une question d'éthique avec les administrateurs systèmes en charge de ces machines [Demarthon et al.-2013].

3.2. Droit à l'oubli

La confidentialité des données et le respect de la vie privée semblent surtout soulever des interrogations. Il en est une qui est pourtant primordiale, c'est celle du droit à l'oubli. Dans ce monde interconnecté où nous ne maitrisons plus nos propres données, quel recours avons-nous ?

Si les textes de loi et les recommandations de la CNIL indiquent clairement qu'on ne peut conserver les données au-delà d'un certain laps de temps, l'usage des données de masse contredit ce même aspect. Nous offrons même ces données sans limitation, car en acceptant les clauses d'utilisation des produits de Google, nous reconnaissons qu'ils ne les supprimeront pas. Certaines sociétés bien sûr proposent des solutions pour aider une personne à effacer ses traces numériques, mais cela est bien souvent assez prohibitif et incomplet, Internet ayant ses propres archives.

Ce droit à l'oubli est pourtant bien intégré à la loi « Informatique et Libertés », mais comme nous l'avons vu, l'hébergement des données à l'étranger complique cette gestion. Ce n'est pas comme si nous voulions réellement nous cacher, ou cacher des choses, a expliqué Julien Vaubourg de Lorraine Data Network lors d'une conférence. C'est que, par exemple, nous savons aujourd'hui que les fichiers de polices contiennent 83 % d'erreurs. Sans correction de ces erreurs, des amalgames peuvent être faits. Il est possible d'affirmer que la persistance des données, leur non-effacement automatique, pose un réel problème, car ouvrant sur une sorte de profilage et donc une violation manifeste de la vie privée [Vaubourg-2014].

Un exemple qui illustre ce souci du droit à l'oubli et du respect de la loi « Informatique et Libertés », est celui de la FNAC, qui durant l'été 2013 a racheté le fichier client de Virgin pour 54 000 €. Bien que ce fichier contienne des données sur des clients, ces derniers n'ont pas été consultés sur leur acceptation, que leurs données soient transmises à un tiers. Et il n'y a pas de réel recours possible pour faire valoir ses droits et ainsi être retiré du dit fichier.

3.3. Un débat ouvert

Ces questions sont donc ouvertes et il semble important d'avoir ce débat. Il est possible de rester sur le discours du « je n'ai rien à cacher », mais cela ne veut pas dire que certaines limites de la vie privée ne seront pas bafouées. Il semble important de préciser les frontières et de faire valoir nos droits avant que la situation ne soit hors de contrôle, si elle ne l'est pas déjà.

Citons Gilles Babinet qui lors d'une interview accordée à [Thinkerview-2014], s'est exprimé sur cette problématique. Entre autres, il a illustré ses propos avec le domaine de la santé, où nous pourrions partir d'un postulat initial basé sur la confiance. Malheureusement les travers sont vite possibles et que dirions-nous si les assureurs, ayant accès à ces données, se mettaient à exclure certaines personnes ou à les surfacturer ?

Si un réel débat démocratique sur ces sujets n'a pas lieu, la prise de conscience sera alors trop tardive, comme le montre la loi de programmation militaire en France par exemple. Cette dernière donne la possibilité, entre autre, aux services de sécurité gouvernementaux (armée, forces policières…) de surveiller les citoyens sur les réseaux informatiques, d'accéder à leurs données, sans l'autorisation d'un juge. Pour faire une comparaison, il s'agit de l'équivalent du « Patriot Act » américain pour la France.

Conclusion

Comme nous l'avons vu, l'exploitation des données de masse offre des débouchés très variés. Il n'existe pas une solution privilégiée ni de biais d'étude recommandé. C'est vraiment le talent du data scientist et la richesse des interactions avec les autres services (métier, informatique…) qui permettra d'aider à la bonne exploitation des données. Nous entrons dans une ère où nous sommes submergés par les informations. À nous de savoir les exploiter sans nous « noyer ». Cela demande autant de rigueur que de créativité et les perspectives semblent encore infinies.

Nous avons pu constater que les données d'un domaine peuvent aider aussi bien leur propre secteur qu'un autre sans rapport apparent. Il en va ainsi d'Abidjan, où l'exploitation des données de téléphonie a permis de projeter les mouvements d'usagers. De même en médecine, l'exploitation des grosses données permet des gains en séquençage génétique ou en épidémiologie.

De nombreux éditeurs se sont lancés dans ces études et chaque semaine des annonces sont faites dans ce secteur. Que ce soit par rapport à de nouvelles solutions ou usages, ou bien par rapport à des vols de données. Les données ont réellement acquis une valeur et leur exploitation attire comme l'Eldorado en son temps.

Il est néanmoins important de considérer plusieurs points qui pourraient refroidir cet engouement. Tout d'abord, étudier la pérennité des données, aussi bien pour leur stockage à venir, que par rapport à leur persistance serait important. De plus, un réel débat devrait avoir lieu sur l'utilisation de ces données, sur les limites à mettre en place (conservation, modifications…). En effet, les dérives par rapport au respect de la vie privée représentent un risque important. Mais un débat est-il encore possible ou bien est-ce déjà trop tard dans une société où les technologies avancent par grands bonds ?

Une autre question qui se pose est le nombre de spécialistes des données disponibles actuellement. Des études tendent à montrer que nous risquons de nous heurter à une pénurie de profils compétents (statisticiens, mathématiciens…) dans les dix ans à venir.

Les données de masse et leur exploitation prennent place dans un contexte plus vaste, comme nous pouvons le voir par l'utilisation de plus en plus intensive de l'informatique ubiquitaire et la réalité chaque jour croissante de l'Internet des objets. Il nous appartiendra en revanche de réussir à placer l'homme au centre de nos réalisations, d'être anthropo-centrés si nous souhaitons réellement accompagner ces succès.

Bibliographie

Alternativ'City. 2013. « Vers une ville intelligente et durable ». Supplément Mag2 Lyon avril 2013. Page 21. Consulté le 17.04.2014

Apache Hadoop. Welcome to Apache Hadoop. Disponible sur : http://hadoop.apache.org/ . Consulté le 25.04.2014.

Braly Jean-Philippe. Décembre 2013. « Yaniv Erlich : Nous avons cassé l'anonymat de données génétiques ». La Recherche n°482. Page 41. Consulté le 14.04.2014.

Brasseur Christophe. Enjeux et usages du Big Data (technologies, méthodes et mise en œuvre). 2013. Editions Lavoisier. 210 pages. Consulté le 30.04.2014

Buchet Jean-Louis. 23.01.2014. « Buenos Aires accélère sa métamorphose en « ville intelligente » ». *La Tribune*. Disponible sur : http://www.latribune.fr/actualites/economie/international/20140122trib000811255/buenos-aires-accelere-sa-metamorphose-en-ville-intelligente.html . Consulté le 12.04.2014.

Calabrese Francesco. Décembre 2013. « Un réseau d'autobus redessiné grâce au téléphone mobile ». La Recherche n°482. Pages 32 à 35. Consulté le 14.04.2014.

Chalmers Sean, Bothorel Cécile, Picot-Clemente Romain. 13.11.2014. Big Data – State of the Art. Disponible sur : http://hal.archives-ouvertes.fr/docs/00/90/39/66/PDF/StateOfTheArt_whichBigDataToolsChoose.pdf . Consulté le 20.04.2014.

Demarthon Fabrice, Delbecq Denis, Fléchet Grégory. Janvier 2013. The Big Data Revolution. *CNRS international magazine n°28*. Disponible sur : http://www2.cnrs.fr/en/2155.htm . Consulté le 17.04.2014.

Féraud Jean-Christophe & Chetrit Judith. 22.09.2013. Toute la ville en smart In Libération. Disponible sur : http://www.liberation.fr/economie/2013/09/22/toute-la-ville-en-smart_933813 . Consulté le 12.04.2014.

Foucret Aurélien. 2011. NoSQL, Une nouvelle approche du stockage et de la manipulation des données In Smile France. Disponible sur : http://www.smile.fr/Smile-france/Livres-blancs/Culture-du-web/Nosql . Consulté le : 21.04.2014.

France Métropolitaine In Réseau Sentinelle. 2013. Disponible sur : http://websenti.u707.jussieu.fr/sentiweb/ . Consulté le 28.04.2014.

GEFORCE. Spécifications techniques GPU GeForce GTX 780. Disponible sur : http://www.geforce.com/hardware/desktop-gpus/geforce-gtx-780/specifications .Consulté le 28.04.2014

Guillaud Hubert. 04.01.2014. En quoi les Big data sont-elles personnelles ? In Blog Les Echos. Disponible sur http://blogs.lesechos.fr/internetactu-net/en-quoi-les-big-data-sont-elles-personnelles-a14016.html . Consulté le 02.05.2014.

INRIA. Un partenariat CITRIS - Inria - PRIME sur les Smart Cities In Inria.fr. Disponible sur : http://www.inria.fr/actualite/actualites-inria/partenariat-citris-inria-prime . Consulté le 13.04.2014.

La NASA gère une avalanche de « Big Data » qui profite à toute la planète In IIP Digital. 2013. Disponible sur : http://iipdigital.usembassy.gov/st/french/article/2013/10/20131021284948.html#axzz30BN3d1IC . Consulté le 28.04.2014.

Laney Doug.06.02.2001. 3D Data Management : Controlling Data Volume, Velocity, and Variety. *META Group – Application Delivery Strategies*. Disponible sur : http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf . Consulté le 05.05.2014.

Linked Data In Tim Berners-Lee, W3.org. Disponible sur : http://www.w3.org/DesignIssues/LinkedData.html . Consulté le 13.04.2014.

Marshall Alex. February 2014. « Big Data, Big Questions ». *Metropolis Magazine*. Disponible sur : http://www.songdo.com/Uploads/FileManager/Metropolis%20Magazine%202.2014.pdf . Consulté le 02.05.2014.

Meldrum Cliff, A Doyle Maria, W Tothill Richard. Novembre 2011. Next-Generation Sequencing for Cancer Diagnostics : a Practical Perspective. Clin Biochem vol 32. Pages 177 à 195. Consulté le 19.04.2014.

Montjoye et al. 25.03.2013. Unique in the Crowd: The privacy bounds of human mobility In Scientific Reports. Disponible sur : http://www.nature.com/srep/2013/130325/srep01376/full/srep01376.html . Consulté le 20.04.2014.

Open Data : 23 jeux de données sur la Recherche et l'Enseignement supérieur In Le Monde Informatique. Bertrand Lemaire. 2014. Disponible sur : http://www.lemondeinformatique.fr/actualites/lire-open-data-23-jeux-de-donnees-sur-la-recherche-et-l-enseignement-superieur-57279.html . Consulté le 24.04.2014.

Picon Antoine. 2013. Smart Cities. Editions B2.120 pages. Consulté le 14.04.2014.

Quand le big data veut changer le quotidien In France 24. Seibt Sébastian. 2014. Disponible sur : http://www.france24.com/fr/20140402-big-data-salon-paris-parcmetre-parkeon-meteo-ericsson-sante-openhealth/ . Consulté le 28.04.2014.

Ranck Jody, 2012. “Connected Health: How Mobile Phones, Cloud, and Big Data Will Reinvent Health Care”. Gigaom Books. Livre électronique. 172 pages. Consulté le 22.04.2014.

SOAP, Short Oligonucleotide Analisys Package In genomics.org. 2007-2010. Disponible sur http://soap.genomics.org.cn/ . Consulté le 17.04.2014.

Special Investigation, Big Data les nouveaux devins. 2013. Produit par KM, par François Lescalier pour Canal +. Consulté le 13.04.2014.

Ten Principles for Opening Up Government Information In Sunlight Foundation. 2010. Disponible sur : http://sunlightfoundation.com/policy/documents/ten-open-data-principles/ . Consulté le 13.04.2014.

Thinkerview. 14.04.2014. Interview Gilles Babinet. Le Big Data. Disponible sur : https://www.youtube.com/watch?v=m0ha0aN1cSU . Consulté le 14.04.2014.

Vaubourg Julien. 17.04.2014. « Je n'ai rien à cacher ». Journée ISN-EPI (Loria). Consulté le 19.04.2014.

Big Data en médecine, en smart cities, … Principes, utilités, exemples et solutions.

Mémoire présenté en vue d'obtenir
UE « Information et communication pour ingénieur »
Spécialité : INFORMATIQUE
Lyon, 2014

RESUME

Le Big Data, ou données de masses est un domaine connaissant une forte croissance ces dernières années. Il est né de l'avalanche croissante de données. Chaque année, il s'en produit plus que les précédentes avec une croissance exponentielle.
Il existe diverses solutions techniques pour stocker et exploiter de tels volumes, mais aussi pour les valoriser, en faisant un réservoir de valeur. Nous pouvons assister à la maturation de nouvelles architectures, mais aussi à la naissance de nouveaux métiers comme le « data scientist ».
Ces données sont maintenant exploitées dans de nombreux domaines, du commerce ou marketing, à la médecine avec l'épidémiologie et le séquençage génétique des génomes, ou dans les villes intelligentes qui se construisent autour de nous, alliant gestion et optimisation de l'énergie, des transports, des communications…
De nombreuses questions restent néanmoins posées sur la propriété et pérennité des données, mais aussi sur le respect de la vie privée.

Mots clés : Big Data, Smart City, smartcities, Ville intelligente, scalabilité, informatique en nuage, données de masse, grosses données, vie privée, scientifique des données, analyste des données

---

SUMMARY

The past years, Big Data became an expending sector. It was born from the growing avalanche of data. In fact, each year we produce more and more data than the previous, with an exponential growth.
There are various technical solutions to store, operate to value such volumes, and making it a value storage. We witnessed the maturation of new architectures, and also the birth of new professions like "data scientist".
From now on, this data is use in many fields from marketing, medicine with epidemiology, genetic sequencing, or in smart cities, built around us, combining management and optimization of energy, transport, communication…
However, many questions remain, about ownership and sustainability of data, and about respect of privacy.

Key words : Big Data, Smart City, smartcities, scalability, Cloud computing, Cloud, privacy, data scientist, data analyst