eDataPrivacy

Réconcilier Big Data et protection de la vie privée

Livre blanc - RGPD et Big Data

Ce livre blanc a été publié alors que j'exerçais pour Umanis.

Patricia.chemali[at]edataprivacy.fr
28/07/2020

Table des matières

I - Introduction

Toujours plus loin, toujours plus vite, toujours plus, la donnée est un marché exponentiel et lucratif. La donnée est le nouvel eldorado lit-on souvent.

La CNIL définit le Big Data comme étant « …Le gigantesque volume de données numériques produites combiné aux capacités sans cesse accrues de stockage et à des outils d'analyse en temps réel de plus en plus sophistiqués. Il offre aujourd'hui des possibilités inégalées d'exploitation des informations. Les ensembles de données traités correspondant à la définition du big data répondent à trois caractéristiques principales : volume, vélocité et variété. »

Au-delà de cet enthousiasme du volume, la variété et de la vélocité et toutes leurs promesses, la réalité du Big Data est bien souvent plus proche du casse-tête de que l'eldorado.

Si le Big Data ouvre des perspectives intéressantes, il n'en présente pas moins certains écueils.

Comment maîtriser dans la durée la croissance du volume de données ? Elles doublent tous les 2 ans.
La question d'une variété de données efficaces, utiles, pertinentes pour le client : le travail préalable de préparation et d'organisation des données est évalué par Oracle à environ 50 à 80% du temps des spécialistes. La technologie du Big Data est en constante évolution ou révolution, la question de la maîtrise de cette technologie est un enjeu continu.[1]

Pour autant le jeu en vaut la chandelle, le Big Data ouvre des perspectives de connaissance jamais imaginée auparavant. Ces énormes volumes de données peuvent être utilisés pour résoudre des problèmes que vous n'auriez jamais pu résoudre auparavant.

Et l'homme dans tout ça ? nous dit le RGPD

Gilles Babinet a demandé à Erik Orsenna de préfacer son ouvrage : BIG DATA Penser l'homme et le monde autrement (Citation d'Erik Orsenna) : *« Big Data ou Big Brother ? Big data et Big Brother ? Il y a quelques temps, la science-fiction nous faisait frémir en annonçant un âge où les ordinateurs domineraient les humains. Aujourd'hui, la menace se précise. Les données ont-elles pris le contrôle de nos vies ? Désormais, nous sommes suivis, pistés, démasqués, mis en catégories, enregistrés. A chacune de nos innombrables connexions quotidiennes, nous dévoilons un peu plus de notre intimité. On sait tout de nous, de nos préférences de nos espoirs, de nos petites manies inavouables…*

Qu'en est-il de nos droits, à commencer par deux d'entre eux, le droit à l'ombre et le droit à l'oubli ? Comment vivre sous ce projecteur perpétuel ? Que fait-on de toutes ces informations qui nous ont été volées ? Et qui les rassemble, et qui décide, un beau jour, une sale nuit, de les utiliser ? »

Ces craintes sont-elles légitimes ou non légitimes, la question n'est pas là. Le terme Big Data est effectivement trop souvent associé à Big Brother dans l'entendement du commun des mortels. Le sens « Brother » prend de plus en plus une connotation de malveillance.

Citons l'article de l'AFP du 17 septembre 2018, il informe de la mise sur étagères d'outils anti-bloqueur de publicités :

«… Inside Secure, une entreprise française qui s'est fait un nom notamment dans la gestion numérique des droits pour de grands noms comme HBO, Sky ou SFR (groupe Altice), va présenter la semaine prochaine de nouveaux outils antiblocage. Ces outils permettent de « brouiller » les messages envoyés par un site internet à l'ordinateur de l'internaute, pour que le bloqueur de pub ne parvienne pas à distinguer le contenu

[1] https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html

publicitaire et à le supprimer, explique à l'AFP Cyrille Ngalle, vice-président chargé de la protection des contenus chez Inside Secure…. » <

Ici encore l'intention du visiteur de site web de voir respecter son espace intime de visite internet, est considéré comme un frein au commerce : il faut contourner l'intention du visiteur pour l'obliger à recevoir malgré lui des messages publicitaires.

N'irions-nous pas vers une crise de confiance entre l'individu et l'entreprise utilisatrice de données personnelles ?

Le Règlement Général sur la Protection des Données personnelles fixe les règles pour toutes les entreprises européennes. Elles sont contraignantes et vécues très souvent comme une atteinte à la liberté d'entreprendre.

Alors doit-on envisager le RGPD comme un frein au Big Data : Privacy by Design versus Big Data ou une réconciliation est-elle possible : Privacy by Design et Big Data réconciliés.

II - Définitions

1) Big Data

a) Qu'est-ce que le Big data ?

Wikipédia nous dit : « Le Big Data /ˌbɪg ˈdeɪtə/, les mégadonnées ou les données massives, désigne les ressources d'informations dont les caractéristiques en termes de volume, de vélocité et de variété imposent l'utilisation de technologies et de méthodes analytiques particulières pour générer de la valeur. »

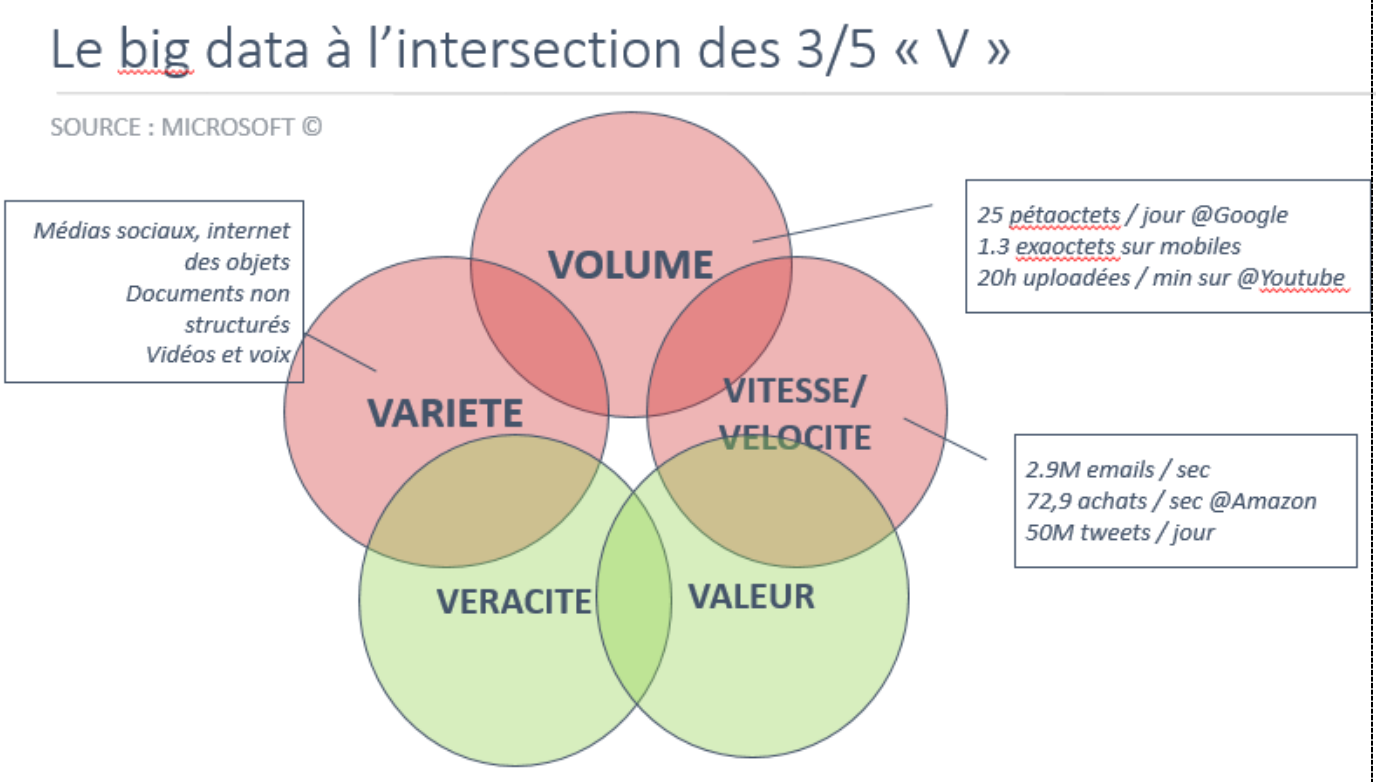
« Littéralement, mégadonnées, grosses données ou encore données massives**.** Big Data désigne un ensemble très volumineux de données qu'aucun outil classique de gestion de base de données ou de gestion de l'information ne peut vraiment travailler. En effet, nous procréons environ 2,5 trillions d'octets de données tous les jours**.** Ce sont les informations provenant de partout **:** messages que nous nous envoyons, vidéos que nous publions, informations climatiques, signaux GPS, enregistrements transactionnels d'achats en ligne et bien d'autres encore. Ces données sont baptisées Big Data ou volumes massifs de données**.** Les géants du Web, au premier rang desquels Yahoo (mais aussi Facebook et Google), ont été les tous premiers à déployer ce type de technologie.
Cependant, aucune définition précise ou universelle ne peut être donnée au Big Data**.** Etant un objet complexe polymorphe, sa définition varie selon les communautés qui s'y intéressent en tant qu'usager ou fournisseur de services. Une approche transdisciplinaire permet d'appréhender le comportement des différents acteurs : les concepteurs et fournisseurs d'outils (les informaticiens), les catégories d'utilisateurs (gestionnaires, responsables d'entreprises, décideurs politiques, chercheurs), les acteurs de la santé et les usagers.[2] »

b) Définition du Big Data: les 3V versus les 5V

Le Big Data se définit historiquement par la réunion des 3 V

- **Volume** : d'énormes quantités de données à l'échelle de zettaoctets et plus. **Le volume correspond à la masse d'informations produite chaque seconde**. Selon des études, pour avoir une idée de l'accroissement exponentiel de la masse de données, on considère que **90 % des données ont été engendrées durant les années où l'usage d'internet et des réseaux sociaux a connu une forte croissance**. L'ensemble de toutes les données produites depuis le début des temps jusqu'à la fin de l'année 2008, conviendrait maintenant à la masse de celles qui sont générées chaque minute. Dans le monde des affaires, le volume de données collecté chaque jour est d'une importance vitale.
- **Vitesse** : **La vitesse ou vélocité équivaut à la rapidité de l'élaboration et du déploiement des nouvelles données**. Par exemple, si on diffuse des messages sur les réseaux sociaux, ils peuvent devenir « viraux » et se répandre en un rien de temps. Il s'agit d'analyser les données au décours de leur lignée (appelé parfois analyse en mémoire) sans qu'il soit indispensable que ces informations soient entreposées dans une base de données. (Flux de données en temps réel provenant de diverses ressources (par exemple, des capteurs physiques ou des « capteurs virtuels » des médias sociaux, tels que les flux Twitter).
- **Variété** : **Seulement 20% des données sont structurées puis stockées** dans des tables de bases de données relationnelles similaire à celles utilisées en gestion comptabilisée. **Les 80% qui restent sont non-structurées**. Cela peut être des images, des vidéos, des textes, des voix, et bien d'autres encore… La technologie Big Data, permet de faire l'analyse, la comparaison, la reconnaissance, le classement des données de différents types comme des conversations ou messages sur les réseaux sociaux, des photos sur différents sites etc. Ce sont les différents éléments qui constituent la variété offerte par le Big Data. (Données provenant d'une vaste gamme de systèmes et de capteurs, dans différents formats et types de données.)

[2] https://www.lebigdata.fr/definition-big-data#:~:text=Le%20ph%C3%A9nom%C3%A8ne%20Big%20Data,et%20la%20pr%C3%A9sentation%20des%20donn%C3%A9es.



Mais progressivement, cette définition a été complétée de 2V : Véracité et Valeur.

- **La véracité :** La véracité concerne la fiabilité et la crédibilité des informations collectées**.** Comme le Big Data permet de collecter un nombre indéfini et plusieurs formes de données, il est difficile de justifier l'authenticité des contenus, si l'on considère les post Twitter avec les abréviations, le langage familier, les hashTag, les coquilles etc. Toutefois, les génies de l'informatique sont en train de développer de nouvelles techniques qui devront permettre de faciliter la gestion de ce type de données notamment par le W3C.
- **La valeur :** La notion de valeur correspond au profit qu'on peut tirer de l'usage du Big Data. Ce sont généralement les entreprises qui commencent à obtenir des avantages incroyables de leur Big Data. Selon les gestionnaires et les économistes, les entreprises qui ne s'intéressent pas sérieusement au Big Data risquent d'être pénalisées et écartées**.** Puisque l'outil existe, ne pas s'en servir conduirait à perdre un privilège concurrentiel.

2) Le Big Data selon le RGPD :

Le RGPD[3] définit la données personnelles comme étant « toute information se rapportant à une personne physique identifiée ou identifiable [...] ; est réputée être une « personne physique identifiable » une personne physique qui peut être identifiée, directement ou indirectement, notamment par référence à un identifiant, tel qu'un nom, un numéro d'identification, des données de localisation, un identifiant en ligne, ou à un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale ». Cette définition est légèrement plus approfondie que celle qui figurait à l'article 2 de la directive 95/46/CE, laquelle précisait que la personne devenait identifiable « notamment par référence à un numéro

[3] *Règlement Général sur la Protection des Données personnelles. Ce règlement européen s'applique à toute entreprise établie en Europe ou utilisant des données personnelles de résidants européens.*

d'identification ou à un ou plusieurs éléments spécifiques, propres à son identité physique, physiologique, psychique, économique, culturelle ou sociale » (Article 4 du RGPD).

3) Notion de Big Data analytics

La finalité essentielle de la collecte en grands volumes, en grandes variétés de données est leur analyse automatisée à grande vitesse pour obtenir les résultats les plus fiables donc de la plus grande valeur, d'où l'introduction de la notion de Big Data Analytics.

Les outils de *Big Data* et les *analytics* sont utilisés dans presque tous les secteurs d'activité. Ils occupent une place de plus en plus importantes dans notre société : le sport de haut niveau, le programme de surveillance PRISM de la NSA**,** la médecine analytique ou encore les algorithmes de *recommandation* d'Amazon (profilage selon le RGPD).

En entreprise particulièrement, l'usage de ces outils permet l'amélioration de l'expérience client, l'optimisation de processus et de la performance opérationnelle, le renforcement ou diversification d'un business model, ou encore une différenciation concurrentielle.
Toutefois, grands volumes et variétés imposent à l'administrateur de prendre en compte à la source les enjeux de :

- gestion rentable des données,
- optimisation du stockage d'informations,
- analyses programmables (à anticiper et mise en place),
- manipulation des données (à simplifier).

III – Big Data versus RGPD

A - Confidentialité et Big Data : quoi de neuf aujourd'hui ?

LES DÉFIS DE CONFIDENTIALITÉ (PRIVACY)

En ce qui concerne les 3/5V, les principaux enjeux de confidentialité dans le Big Data sont :

Manque de contrôle et de transparence (2 principes fondamentaux de conformité au RGPD): les sources de données dans le Big Data sont multiples, parfois inattendues ce qui peut compliquer voire rendre impossible le contrôle par un individu sur ses données personnelles. Dans de nombreux cas, il n'est même pas au courant du traitement ou ne peut pas suivre la façon dont les données circulent d'un système à un autre. Or un des principes fondamentaux du RGPD est celui du contrôle par le responsable du traitement sur les données personnelles pour permettre à l'individu d'avoir un contrôle sur ses informations.

Défi 1 : La gestion / politique du consentement est un enjeu.

Réutilisation des données : l'évolutivité du stockage permet d'envisager un espace infini, ce qui signifie que les données peuvent être collectées en continu jusqu'à ce qu'une nouvelle valeur puisse être créée à partir des informations dérivées.

Défi 2 : la politique de rétention et la transparence incluent ces problèmes

Inférence et ré-identification des données : Le choix est souvent fait de contourner les enjeux de la conformité RGPD en optant pour la mise en place d'un processus d'anonymisation. Or ce contournement n'est efficace

que aucune ré identification n'est possible ce qui est souvent handicapant pour les utilisations métiers des données du Big Data. La ré-identification est possible en combinant diverses données prétendument non personnelles pour déduire des informations relatives à une personne ou à un groupe (un triplet de données non personnelles permet une ré-identification – démonstration Sweeney de 2002 – voir définition Données Anonymisées). Une combinaison d'ensembles de données anonymes et d'analyses avancées peut conduire à la ré-identification d'une personne en extrayant et en combinant différentes informations.

Défi 3 : Organiser un système continu d'enquêtes sur la confidentialité

Profilage et prise de décision automatisée : les analyses appliquées à des ensembles de données combinés visent à créer des profils spécifiques pour les individus qui peuvent être utilisés dans le contexte de systèmes de prise de décision automatisés, par ex. pour offrir ou exclure des services et produits spécifiques. Un tel profilage peut dans certains cas conduire à un isolement et / ou à une discrimination, y compris une différenciation des prix, sans donner aux individus la possibilité de contester ces décisions. De plus, le profilage, lorsqu'il est basé sur des données incomplètes, peut conduire à de faux négatifs, privant injustement les individus des droits et avantages auxquels ils ont droit. Le RGPD prévoit que tout individu doit pouvoir s'opposer à un traitement automatisé de ses données personnelles.

Défi 4 : la difficulté de faire appliquer et / ou de surveiller les contrôles de protection des données. (// Cloud SaaS)

Par nature, le Cloud augmente la surface d'exposition aux menaces et donne plus de fenêtres d'opportunités aux attaquants. La dispersion des données est un enjeu : difficile de protéger chaque donnée quand celles-ci sont éparpillées dans les serveurs d'une architecture multi-Cloud.
Toutefois la force de frappe technologique des fournisseurs de Cloud et la précision des services de détection mise en place, le Cloud peut sembler dans certains cas plus à même de répondre à des attaques de grande envergure avec une réaction sur des délais extrêmement brefs qu'un hébergeur, mainteneur de solutions ne sera pas en mesure de mettre en œuvre.

Défi 5 : Partager les rôles et responsabilités avec l'ensemble des parties prenantes

L'implication et l'interaction de nombreuses parties prenantes diverses, complique (pour les régulateurs, les contrôleurs de données et les utilisateurs) l'identification des failles de confidentialité et la mise en place de mesures de sécurité adéquates. Le RGPD considère qu'un traitement est légitime si des mesures de sécurité adéquates ont été mises en place.

B – Que serait le Big Data sans le RGPD ?

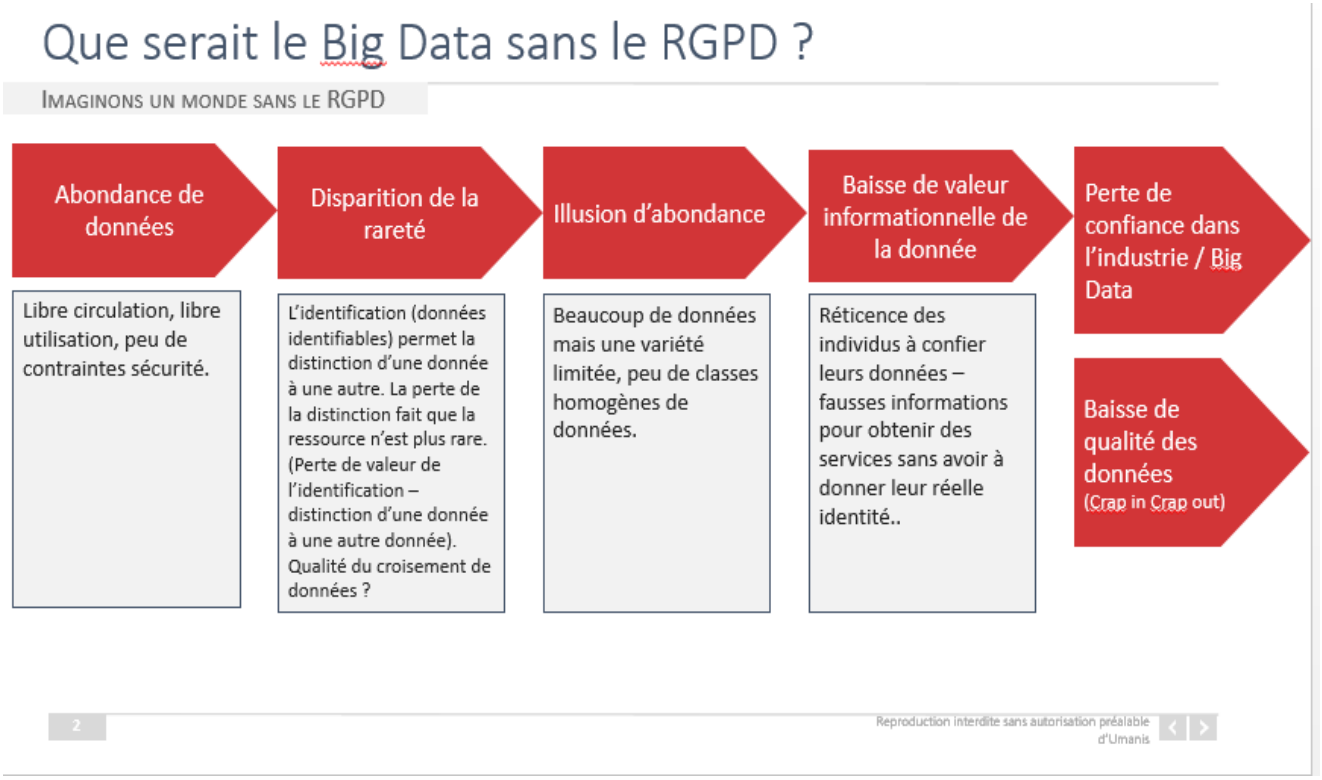
ENISA – 2015 : « Personne ne tirera profit d'un clash entre Big Data et Privacy »

Source : ENISA Report 2015 « *L'oxymore du Big Data et de la confidentialité*

Imaginons un monde sans GDPR : dans un tel scénario, les données se référant à des personnes identifiées ou identifiables, dans la mesure où elles se distinguent les unes des autres en termes d'attributs, ne seraient plus une ressource rare.
Mais si les données personnelles étaient aussi largement disponibles, et leur variété à part entière contenue dans un nombre, même si apparemment large, mais néanmoins limité de classes homogènes, sans aucune forme de protection réglementaire de leur rareté, leur valeur informationnelle serait moindre. Par exemple, les gens commenceraient progressivement à être plus réticents à fournir leurs données ou ils fourniraient de fausses données, afin d'obtenir les services qu'ils souhaitent sans découvrir leur identité.
Cela ouvrirait la voie à une grave réduction de la qualité des données. »

L'ENISA encourage ainsi l'industrie du Big Data à envisager le RGPD sous l'angle de la protection de la rareté de la Donnée, car la rareté des données personnelles doit être considérée comme une valeur à protéger.
D'autres auteurs évoquent la lassitude, « …l'exaspération qu'un sur-usage des données peut engendrer auprès d'un prospect ou d'un client »[4].

Privacy - Construire la confiance dans le Big Data ET confiance dans l'industrie : personne ne profitera d'une rupture de cette confiance. Si les utilisateurs sentent que leurs données personnelles ne sont pas correctement protégées, ils s'orienteront vers des solutions de contournement (quel que soit le temps que cela pourra prendre). Un exemple à l'appui de cette théorie est celui de la publicité comportementale basée sur les cookies. Les agences et réseaux de publicité n'ayant pas réussi à adopter des mécanismes de consentement appropriés, les utilisateurs se sont lentement dirigés vers l'adoption généralisée des logiciels bloqueurs de publicités (qui ont en fait un impact beaucoup plus grand sur l'industrie de la publicité – Voir article AFP du 7 sept 2018 en annexe).



[4] Big Data et Machine learning, les concepts et les outils de la data science : Pirmin Lemberger, Marc Batty, Médéric Morel, Jean-luc Raffaëlli

IV - Privacy by Design et Big Data réconciliés

1) Privacy by Design dans le Big Data

Ou, *Mécanisme pour aborder les risques de respect de la vie privée ou Privacy, dès le début du traitement de données et appliquer les solutions de préservation de la vie privée nécessaires à toutes les étapes de la chaîne de valeur du Big Data.*

Privacy by Design : concept et stratégies de conception

Un concept à multiples facettes : la protection de la vie privée dès la conception n'est ni un ensemble de simples principes généraux ni ne peut être réduite à la mise en œuvre de Privacy Enhancing Technology[5]. Il s'agit d'un processus impliquant divers composants technologiques et organisationnels, qui mettent en œuvre les principes de confidentialité et de protection des données.
8 principes suggérés :

PRIVACY BY DESIGN : STRATEGIE	DESCRIPTION	Les enjeux du Big Data
Minimiser	Le nombre de données personnelles utilise doit être réduit au minimum – le plus petit possible (data minimization).	Mettre en place une collecte plus utile et qualitative.
Masquer	La donnée personnelle et ses inter – relations doivent être masquées à l'écran.	Ne devrait pas soulever de défi particulier
Cloisonner	Les données personnelles doivent être traitées de manière distribuée, dans des espaces dédiés et séparés chaque fois que possible.	Ne devrait pas soulever de défi particulier
Agréger	Les données personnelles doivent être traitées au niveau d'agrégation le plus élevé et avec le moins de détails possibles dans lesquelles elles sont (encore) utiles.	Ne devrait pas soulever de défi particulier
Informer	Les personnes concernées doivent être informées de manière adéquate de toutes les utilisations qui sont faites de leurs données (transparence).	Mettre en place des mécanismes d'information et transparence à destination des personnes concernées.
Contrôler	Les personnes concernées devraient être informées des conditions du traitement de leurs données personnelles.	Concevoir un process + outil de collecte de consentement et préférences + gestion des STOP & GO -
Renforcer	Une politique de confidentialité compatible avec les exigences légales doit être en place et doit être appliquée.	Tous les RT doivent mettre en place et appliquer leurs politiques de

[5] Privacy Enhancing Technology ou PET : *Une technologie de protection de la vie privée (PET) est une méthode de protection des données. Les PET permettent aux utilisateurs en ligne de protéger la confidentialité de leurs informations personnelles identifiables (PII) fournies et traitées par des services ou des applications. Les PET utilisent des techniques pour minimiser la possession de données personnelles sans perdre les fonctionnalités d'un système d'information.*

		confidentialité, conformément au principe de responsabilité
Démontrer	Les RT doivent être en mesure de démontrer le respect de la politique de confidentialité en vigueur et de toutes les exigences légales applicables.	Ne devrait pas soulever de défi particulier

Privacy by Design et Big Data Analytics

BIG DATA CHAINE DE VALEUR	LES PRINCIPES DU PRIVACY BY DESIGN	MISE EN ŒUVRE
ACQUISITION / COLLECTE DE DONNEES	MINIMISER	Définir les données strictement nécessaires aux usages et objectifs poursuivis avant la collecte, sélectionner avant la collecte (réduire les champs de données, définir les contrôles pertinents, supprimer les informations indésirables, etc.), conduire les DPIA ou AIPD. Big data et Machine Learning – Pirmin Lemberger : "…l'approche "tout stocker" n'a pas d'avenir : il faudra, tôt ou tard, savoir définir l'obsolescence de ces données tout comme il sera nécessaire de savoir piloter l'assainissement de son patrimoine data.
	AGREGER	Dépersonnalisation / Anonymisation local (Préférentiellement à la source).
	MASQUER	Sélectionner et mettre en place des outils améliorant le respect de la vie privée / confidentialité des personnes concernées, par ex. outils anti-tracking, outils de cryptage, outils de masquage d'identité, partage de fichiers sécurisé, etc. (PETs)
	INFORMER	Mettre en place les mécanismes de respect de l'obligation de transparence. Informer les personnes concernées sur toutes les utilisations faites de leurs données et conditions d'utilisation (lister tous les destinataires de données…)
	CONTROLER	Mettre en place un mécanisme d'expression et de collecte de consentement : mécanismes d'Opt-out, mécanismes appropriés d'expression des préférences de sollicitations, '*sticky*' politiques[6],…
ANALYSE DE DONNEES & CONSERVATION / RETENTION	AGREGER	Techniques d'anonymisation reconnues et validées (attention notion d'IRREVERSIBILITE – y compris sur Triplet (k-anonymity family, differential privacy). (voir annexe Définitions)
	MASQUER	Chiffrement interrogeable, préservant le secret et autorisant les calculs.
STOCKAGE DE DONNEES	MASQUER	Chiffrement de toutes les données dormantes. Mécanismes de contrôles et d'authentification. Toutes autres mesures de nature à sécuriser le stockage de données.
	CLOISONNER	Installations de stockage et d'analyse distribuées / décentralisées.
UTILISATION DE DONNEES	AGREGER	Techniques de dépersonnalisation / Anonymisation. Qualité et origine des données.
TOUTES LES PHASES	DOCUMENTER / DEMONTRER	Outils automatisés de définition, d'application, de responsabilité et de conformité de politique.

Engager les mesures de sécurité, pas un problème de budget mais de volonté[7]

[6] https://documents.epfl.ch/users/a/ay/ayday/www/mini_project/Sticky%20Policies.pdf
[7] https://www.federation-eben.com/wp-content/uploads/2017/05/zdnet.fr_14032017_Big-Data_mais-Big-menaces-%C3%A9galement.pdf

Big Data Analytics :

Le terme « Big Data analytics » fait référence à l'ensemble du cycle de vie de la gestion des données consistant à collecter, organiser et analyser des données pour découvrir des modèles, déduire des situations ou des états, prédire et comprendre les comportements.

Sa chaîne de valeur comprend :
Acquisition / collecte de données. Cette étape englobe les processus de collecte, de filtrage et de nettoyage des données avant qu'elles ne soient placées dans un référentiel de données.
Le processus est généralement basé sur une collecte de données rapide et massive, supposant ainsi des données en grand volume, à grande vitesse, à grande variété et à haute véracité mais potentiellement à faible valeur.

Analyse des données. Cette étape consiste à convertir les données « brutes » collectées en données utilisables pour des utilisations spécifiques aux métiers (ie. Aide à la prise de décision…). L'analyse des données porte tant sur les données structurées que non structurées, avec / sans informations sémantiques et peut avoir plusieurs niveaux de traitement et différentes techniques (par exemple, analyse diagnostique, descriptive, prédictive et prescriptive).

Data Curation = contenu : création, sélection, classification, transformation, validation et préservation (la réutilisabilité est un problème).

Stockage de données : stocker et gérer les données de manière évolutive répondant aux besoins des applications / analyses nécessitant un accès aux données.

Utilisation des données : couvre l'utilisation des données par les parties intéressées et dépend beaucoup du scénario de traitement des données.

2) Concevoir des stratégies dans la chaîne de valeur de l'analyse du Big Data

Concevoir un plan d'action de conformité prenant en compte le cycle de vie complet de l'analyse.

a) Acquisition / collecte de données

a. Minimiser
   i. AGREGAT : supprimez toutes les informations personnelles avant de publier les données à des fins d'analyse
   ii. Masquer: voir le marché des PET (anti-tracking, cryptage, masquage d'identité et outils de partage de fichiers sécurisés)
      1. AVIS: donnez des informations adéquates
      2. CONTRÔLE: Opt-in Opt-out + tout outil qui devrait donner le contrôle aux personnes concernées

b) Analyse et conservation des données

b. AGREGER : K-anonymat et confidentialité différentielle sont les deux principales familles de modèles de confidentialité avec différents types d'implémentations. Suivez l'efficacité de l'anonymisation pendant tout le cycle de vie des données. Assurez-vous que votre DPO valide le processus d'anonymisation. (Banques : PCI//RGPD voir Tokénisation//Anonymisation - Pseudonymisation)

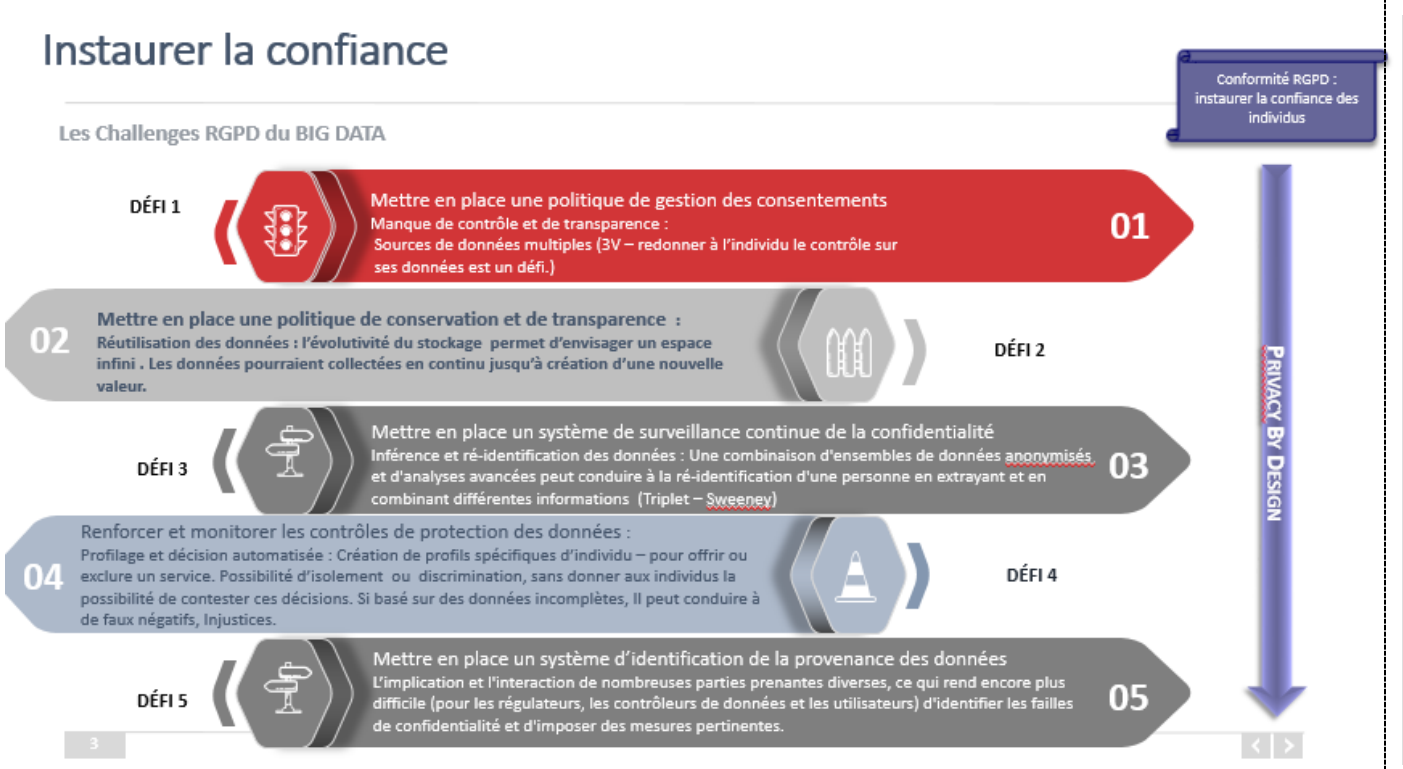
c. MASQUER : Le cryptage interrogeable, le cryptage homomorphique et les calculs multipartites sécurisés sont des technologies prometteuses dans ce domaine avec beaucoup d'intérêt pour la communauté de recherche.

c) Stockage de données

d. MASQUER : un contrôle d'accès et une authentification granulaires sont essentiels pour protéger les données personnelles dans les bases de données. Les technologies telles que le contrôle d'accès basé sur les attributs peuvent être beaucoup plus évolutives dans le Big Data, offrant des politiques de contrôle d'accès précises. Le chiffrement est également essentiel pour protéger les données au repos.

e. CLOISONNER : Les mesures de contrôle d'accès et les techniques de chiffrement peuvent à nouveau prendre en charge ce type de solutions.

3) Les enjeux RGPD du Big Data



V - Recommandations des agences européennes et nationales

Les défis de la technologie pour les Big Data devraient être abordés par les opportunités de la technologie pour la vie privée.

Passer de la discussion « Big Data versus Privacy » à « Big Data et Privacy » signifie que le concept de confidentialité dès la conception est essentiel pour identifier les exigences de confidentialité dès le début de la chaîne de valeur de l'analyse du Big Data.

ENISA : Considérer la confidentialité dès la conception comme un outil essentiel pour faire face aux risques liés au Big Data.

1) Mettre en œuvre une gouvernance - qualité de la donnée :

L'ENISA estime que cet objectif peut être réalisable si toutes les parties prenantes adoptent les mesures nécessaires pour intégrer les garanties du Privacy by Design & by default (confidentialité et protection) au cœur du Big Data. À cette fin, l'ENISA formule les recommandations suivantes:

1. Mettre en place et monitorer les principes du Privacy by Design ;
2. Analyse de données décentralisée versus centralisée (éviter le bruit – privilégier la sélectivité et l'effectivité de la donnée au volume). Mettre en place un processus de sélection des données en fonction de leur nécessité ;
3. Assistance et automatisation de l'application des politiques dans la chaîne du co-contrôle du Big Data et d'échange d'informations : certaines exigences de confidentialité d'un responsable du traitement peuvent ne pas être acceptables ou respectées par un autre. De la même manière, les préférences en matière de confidentialité des personnes concernées peuvent également être négligées ou mal prises en compte. Par conséquent, il est nécessaire de définir et de mettre en œuvre des politiques automatisées, de sorte qu'une partie prenante ne puisse refuser d'honorer la politique d'une autre partie prenante dans la chaîne d'analyse des Big Data. La sémantique et les normes pertinentes, ainsi que les règles appliquées par cryptographie, sont des domaines qui nécessitent une étude approfondie à cet égard. (Un des principes fondamentaux du Privacy by Design est la « Somme non nulle »)
4. Transparence et contrôle : le mécanisme traditionnel de collecte de consentement et de partage des avis de confidentialité n'est pas efficace pour les Big Data. De nouveaux mécanismes doivent être créés : icônes de confidentialité, politiques collantes et entrepôts de données personnelles. L'industrie de l'analyse des Big Data et les responsables du traitement doivent travailler sur de nouvelles mesures de transparence et de contrôle, en mettant les individus en charge du traitement de leurs données (privilégier la gestion autonome via l'espace personnel).
5. Sensibilisation des utilisateurs et promotion des PET : il existe déjà de nombreux outils d'amélioration de la confidentialité pour la protection en ligne et mobile, tels que l'anti-tracking, le cryptage, le partage de fichiers sécurisé et les outils de communication sécurisés, qui pourraient offrir un soutien précieux pour éviter le traitement indésirable des données personnelles. Il faut toutefois encore évaluer la fiabilité de ces outils et leur applicabilité au grand public. Il faut traiter de manière adéquate les aspects liés à la fiabilité et à la convivialité des PET en ligne.
6. Une approche cohérente de la vie privée et des Big Data : les décideurs politiques doivent aborder les principes (et technologies) de la confidentialité et de la protection des données comme un aspect central des projets de Big Data et des processus décisionnels pertinents.

ANSSI

Le RGPD, mais aussi l'ANSSI, CLUSIF, la CNIL recommande la mise en place de mesures de sécurité adéquates pour l'ensemble des données personnelles collectées par l'entreprise. La notion de mesures adéquates laisse entendre que les risques sont identifiés, classifiés et la mesure de sécurité adéquate choisie offre une réduction significative de ce risque. Le RGPD impose ici la conduite d'une Analyse d'impact sur la vie privée.

L'adéquation sécurité v/ risques est mesurée sur les référentiels et bonnes pratiques suivants :

- ✓ ISO/IEC 27001
- ✓ ISO/IEC 27002
- ✓ ISO/IEC 27005
- ✓ ISO/IEC 27018
- ✓ ISO/IEC 27710

- ✓ ISO/IEC 29100
- ✓ PCI DSS
- ✓ HIPAA
- ✓ …

d. Code de bonnes pratiques pour le management de la sécurité de l'information (27002)

- La norme ISO/IEC 27002:2013 constitue un code de bonnes pratiques. Elle est composée de 114 mesures de sécurité réparties en 14 chapitres couvrant les domaines organisationnels et techniques ci-contre.

- C'est en adressant l'ensemble de ces domaines que l'on peut avoir une approche globale de la sécurité des S.I.

CONCLUSION :

D'ici quelques années, le marché du Big Data va se mesurer en centaines de milliards de dollars. D'après le calcul effectué par le cabinet V. Bourne, dans le monde, l'ensemble des dépenses consacrées au Big data, dans les budgets IT des grandes entreprises, représentait un quart du budget total IT en 2018**.** Le Cap Gemini a aussi commandité en 2015 une étude qui a montré que 61% des entreprises sont conscientes de l'utilité du Big Data en tant que "moteur de croissance à part entière"**.** De ce fait, on lui accorde beaucoup plus d'importance que leurs produits et services existants. Cette même étude a encore indiqué que 43% d'entre elles se sont déjà réorganisées ou se restructurent présentement pour exploiter le potentiel du Big Data.

Le RGPD en germination depuis 2012 est vécu comme une contrainte, son utilité n'est pas toujours perçue par les entreprises qui le voient principalement comme une dépense supplémentaire.
La réconciliation est nécessaire.
Les agences sont unanimes à recommander la mise en place des principes du Privacy by Design, qui outre la fiabilité, l'exhaustivité, l'exactitude, l'intégrité, la cohérence, la fraîcheur des données, permet de réinstaurer la confiance des individus et donc de créer ou conserver la valeur.

Communiquer sur la conformité de ses traitements équivaut à créer de la confiance. La prochaine étape du big data sera probablement l'investissement sur la confiance des individus.

ANNEXE 1 : DEFINITIONS

API : Acronyme d'Applications Programming Interface. Prise de courant numérique, l'API est une interface de programmation qui permet à de multiples services numériques de se brancher sur une application pour échanger des données. Elle est généralement ouverte et proposée par l'éditeur de la solution. Elle peut être normalisée par des organismes comme l'ISO ou l'IUT.

Cluster : Grappe de serveurs ou « ferme de données », structure générique des applications distribuées en Big data.

CNIL : Acronyme de Commission Nationale de l'Informatique et des Libertés. La CNIL représente une autorité administrative indépendante chargée de veiller à ce que l'informatique soit au service du citoyen et qu'elle ne porte atteinte ni à l'identité humaine, ni aux droits de l'homme, ni à la vie privée, ni aux libertés individuelles ou publiques.

CRM : Acronyme de Customer Relationship Management. Il s'agit des progiciels qui permettent de traiter directement avec le client, que ce soit au niveau de la vente, du markéting ou des services annexes, et que l'on regroupe souvent sous le terme de front-office, par opposition aux outils de back-office que sont les progiciels de gestion intégrés ou ERP.

Donnée anonymisée : Des données ayant été l'objet d'un procédé d'anonymisation ne sont plus considérées comme des données à caractère personnel. Selon le considérant 26 du règlement général sur la protection des données, « il n'y a […] pas lieu d'appliquer les principes relatifs à la protection des données aux informations anonymes, à savoir les informations ne concernant pas une personne physique identifiée ou identifiable, ni aux données à caractère personnel rendues anonymes de telle manière que la personne concernée ne soit pas ou plus identifiable. Le présent règlement ne s'applique, par conséquent, pas au traitement de telles informations anonymes, y compris à des fins statistiques ou de recherche. »

Néanmoins, il faut différencier les données « réellement anonymisées » et les données « pseudo anonymisées » (utilisant un code de référence « confidentiel ») qui sont fréquemment utilisées dans des domaines comme la recherche médicale. Les données « pseudo anonymisées » (ou codées) restent des données personnelles et sont donc dans le cadre de la réglementation des données personnelles.

En revanche, la collecte de données à caractère personnel, même si celles-ci sont immédiatement anonymisées, reste un traitement de données soumis aux principes de la protection des données.

L'efficacité des techniques d'anonymisation est cependant mise en doute par certains chercheurs. Des données médicales anonymisées (visites d'hôpitaux, consultations médicales) d'employés de l'État du Massachusetts ont été « ré-identifées » en les croisant avec les listes électorales d'une même ville[8]. Le gouverneur même de l'État a pu être ré-identifié, seules six personnes partageant la même date de naissance dont trois de sexe masculin, et, parmi ces dernières, une seule partageait le même code postal, sur un total de 54 000 résidents et sept codes postaux. L'efficacité de l'anonymisation est aussi fonction de la granularité de l'information, car sur des petits échantillons ou dans le cas de granularité très fine allant jusqu'à quelques individus, l'anonymisation devient inexistante.

Plus récemment, une étude de chercheurs du MIT a montré que quatre points géolocalisés étaient suffisants pour identifier 95 % des individus dans une base de données téléphoniques de 1,5 million de personnes.

Donnée personnelle : Des données ayant été l'objet d'un procédé d'anonymisation ne sont pas considérées comme des données à caractère personnel. Selon le considérant 26 du règlement général sur la protection des données, « il n'y a […] pas lieu d'appliquer les principes relatifs à la protection des données aux informations anonymes, à savoir les informations ne concernant pas une personne physique identifiée ou

[8] Démonstration Sweeney de 2002, un triplet suffit à réidentifier un individu, dans l'affaire le triplet était composé d'une date + ZipCode + Sexe.

identifiable, ni aux données à caractère personnel rendues anonymes de telle manière que la personne concernée ne soit pas ou plus identifiable. Le présent règlement ne s'applique, par conséquent, pas au traitement de telles informations anonymes, y compris à des fins statistiques ou de recherche. »

DPI : acronyme de Deep Packet Inspection. Technologie permettant de reconstituer des messages à partir de l'interception des packets IP qui transitent dans un câble de télécommunication sous-marin par exemple. En informatique la Deep Packet Inspection est, pour un équipement d'infrastructure de réseau, l'analyse du contenu (au-delà de l'en-tête) d'un paquet de réseau (IP le plus souvent) de façon à en tirer des statistiques, à filtrer ceux-ci ou à détecter les intrusions, du spam ou tout autre contenu prédéfini. Le DPI peut servir notamment à la censure sur internet ou dans le cadre de dispositifs de protection de la propriété intellectuelle.

ETL : acronyme d'Extract-Transform and Load. Représente une passerelle faite sur mesure entre deux systèmes d'information. Elle est fondée sur des connecteurs (extract) servant à exporter ou à importer les données dans les applications (par exemple des connecteurs Oracle ou SAP…), des transformateurs (transform) qui reformatent les données (agrégations, filtres, conversion…), et des mises en correspondance (load). Les solutions d'ETL sont apparues dès les années 1970 pour faciliter la conversion régulière de données entre applications dans le monde bancaire et financier. Les ETL sont souvent le cauchemar des directeurs informatiques pour leur complexité. Elles s'opposent aux APIs et visent à normaliser les échanges là où les ETL font des traitements spécifiques.

GAFA : l'acronyme de Google, Apple, Facebook et Amazon, quatre grandes firmes américaines emblématiques de ce qu'est l'économie numérique avec son développement très rapide, et dominant chacune leurs marchés.

Hadoop : à la base, Hadoop représente un framework conçu en open source et permettant de réaliser des traitements sur des volumes de données massifs, de l'ordre de plusieurs pétaoctets. Aujourd'hui, il s'agit davantage d'une définition générique d'outils de Big Data open source, compatibles entre eux.

Depuis son apparition au milieu des années 2000 et le projet Apache qui l'a fait naître, Hadoop a connu beaucoup d'évolutions de chacun de ses composants. Mais le principal changement, on le doit à la version 2 en 2013 et l'apparition de YARN (Yet Another Resource Negotiator) qui lui permet d'exécuter d'autres types d'applications que le seul traitement batch de MapReduce : dépassant le modèle stockage / traitement distribué, l'écosystème Hadoop découvre Spark, Storm, le streaming et d'autres enjeux comme la gouvernance ou la sécurité. Avant le prochain big bang de la data locality (cf plus bas) ?

Hana : acronyme de High Performance Analytic Appliance. C'est une technologie in memory de traitement en mémoire de masse, propriétaire, développée par SAP AG. Hana fonctionne en mode massivement parallèle, exploitant ainsi un maximum de processeurs multicoeurs et permettant l'exécution particulièrement rapide des requêtes.

In memory : les systèmes de Big data « en mémoire » travaillent sur des données stockées dans de la mémoire vive (ou flash) pour accélérer le traitement de requêtes nécessitant de faire de nombreux appels de données. Actuellement beaucoup plus coûteux à exploiter que les systèmes en disques traditionnels, ils n'en représentent pas moins une avancée significative par la vitesse de traitements des algorithmes complexes qu'ils permettent.

Learning machine : système qui pratique l'analyse de situations à partir de données et qui est capable de commencer une action en fonction des typologies de données. Par exemple, prévenir un usager d'un dysfonctionnement lorsque l'infrastructure de son opérateur est tombée en panne.

MapReduce : architecture de développement informatique, inventée par Google, dans laquelle sont effectués des calculs parallèles, et souvent distribués, de données potentiellement très volumineuses, typiquement supérieures en taille à 1 téraoctet. Les termes map et reduce ainsi que les concepts sous-jacents sont empruntés aux modèles de programmation. MapReduce permet de manipuler de grandes quantités de données en organisant leur distribution dans un cluster de machines afin d'y être traitées. Ce modèle connaît

un grand succès auprès de sociétés possédant d'importantes quantités de données à traiter, comme Amazon ou Facebook.

Si MapReduce est à la base le nom d'un paradigme de programmation - celui du traitement distribué - il est devenu par extension le nom d'une API (l'API Java MapReduce d'Hadoop). Et si l'API MapReduce historique est moins utilisée (car moins rapide que Spark ou Storm) et a été remplacée par TEZ dans la plupart des composants incontournables d'Hadoop (Hive, PIG, etc.), le modèle de programmation, lui, perdure : on le retrouve d'ailleurs dans Spark et dans la plupart des outils de l'écosystème Big Data. Quant à l'application historique MapReduce, il ne faut pas l'enterrer non plus : pour des volumes de données massifs, elle reste toujours plus efficace que Spark dont le modèle en mémoire sature plus vite…[9]

NoSQL : en informatique, NoSQL (Not Only SQL en anglais) désigne les systèmes de gestion de base de données (SGBD) qui ne sont pas fondés sur l'architecture classique des bases relationnelles. L'unité logique n'y est plus la table, et les données ne sont en général pas manipulées avec le système générique SQL. Les systèmes NoSQL sont généralement rudimentaires en termes de fonctionnalités mais ils permettent une grande agilité dans le traitement des données. Ils se sont progressivement imposés avec l'explosion de données qu'ont constatée les grands acteurs de l'Internet qui ne parvenaient plus à faire fonctionner leurs services avec des systèmes relationnels traditionnels.

Open data : une donnée « ouverte » est une donnée numérique d'origine publique ou privée, accessible (lisible) à tous. Elle peut être produite par une administration, ou une entreprise. Elle est diffusée de manière structurée selon une méthodologie et une licence ouverte garantissant son libre accès et sa réutilisation par tous, sans restriction technique, juridique ou financière.

SaaS : acronyme de Software as a Service. Service logiciel auquel on accède en ligne et que l'on paye en fonction de l'utilisation (par mois, par volume, etc).

Spark : dispositif open source de type Hadoop qui s'affranchit des architectures de type MapReduce en faisant ses traitements directement dans la mémoire vive (in memory). Accélérant ainsi ses traitements de façon importante, allant jusqu'à 100 fois plus vite que les systèmes Hadoop traditionnels, Spark est un programme prioritaire d'Apache Foundation.

Vie privée, données publiques : Nos sociétés assistent, avec Internet et les médias sociaux, à un élargissement de notre vision de la vie privée. Nous partageons tous les jours de nombreuses informations et données, parfois consciemment sur nos profils Facebook, Instagram ou Twitter, parfois involontairement par les traces laissées via l'acceptation de cookies sur certains sites. En publiant et partageant des informations sur nos profils numériques, on accepte de partager nos données avec notre réseau dans u premier temps (amis, *followers*...), à la plateforme ensuite et enfin à toutes les personnes ayant accès à ces profils. Ces informations, même partagées avec le monde entier, restent des données personnelles. Ces nouvelles formes de dévoilement stratégique d'informations personnelles à des fins de gestion du capital social en ligne n'est en rien une renonciation à la *privacy*; notre besoin de protéger notre intimité existe toujours[22]. Il existe de multiples motivations de la révélation de soi sur les réseaux sociaux. En fonction du réseau social observé, plusieurs façons de gérer son capital social apparaissent. Tous permettent un ajustement de la présentation en ligne de l'utilisateur, certains permettent la mise en commun de détails sélectionnés à différentes sphères plus ou moins intimes. Selon le sociologue Antonio Casilli, sur les médias sociaux, la notion de capital social désigne « l'acquisition, via des relations médiatisées pas les TIC, de ressources matérielles, informationnelles ou émotionnelles »[22]. Ce dévoilement et cette gestion du capital social sont soumis à des coûts, comme celle de la perte de *privacy* ; se faire connaître oblige, notamment, à sacrifier une partie de sa vie privée afin d'attirer des connexions[23]. Toujours d'après Antonio Casilli, « aucune des données partagées n'est privée ou publique, elle représente en quelque sorte un signal que les usagers envoient à leur environnement (ici, les membres de leurs réseaux personnels en ligne), afin de recevoir un retour (*feedback*) dudit environnement »[22]. La *privacy* est alors basée sur la recherche d'un accord entre plusieurs parties ; les acteurs sont prêts à confronter

[9] https://www.bigdataparis.com/documents/2020/BDG19_BD_19206.pdf

leurs intérêts et à faire des concessions mutuelles en termes de dévoilement d'informations potentiellement intimes. La confidentialité, l'intimité et la *privacy* ne dépendent pas uniquement de caractères propres à l'individu mais deviennent contextuelles et donc sujettes à concertation collective[22].

ANNEXE 2 – Article AFP du 7 septembre 2018 : Face aux bloqueurs de publicité, la résistance s'organise.

AFP le 7 septembre 2018

FACE AUX BLOQUEURS DE PUBLICITE, LA RESISTANCE S'ORGANISE

« *Les logiciels bloqueurs de publicité continuent de progresser en France, au grand dam des acteurs de la publicité numérique qui voient disparaitre une partie du public qu'ils sont censés toucher. « C'est un peu l'omerta, on parle beaucoup trop peu des bloqueurs par rapport à leur impact » réel, estime auprès de l'AFP Emmanuel Brunet, le patron de la startup Eulerian Technologies dont le métier est justement de vérifier l'impact des campagnes publicitaires numériques. Selon lui, certaines catégories du public deviennent particulièrement difficiles à toucher par la publicité sur la toile, comme les hommes jeunes, actifs et urbains, grands utilisateurs de bloqueurs de pub.*

Ces publics sont particulièrement enclins à utiliser les bloqueurs de publicité, car ils fréquentent des sites où ils sont indispensables, vu leur degré de pollution publicitaire : les sites de téléchargement illégaux, les sites de streaming… « Il faut parfois réorganiser des plans médias pour toucher » ces catégories, en augmentant le recours à des médias « offline » comme la presse écrite ou la télévision, explique Emmanuel Brunet.

Plus d'un cinquième des internautes français utiliseront un bloqueur cette année

Eulerian Technologies a comptabilisé entre 9 et 41% d'internautes équipés d'un bloqueur de publicité sur les sites que la société étudie, avec des taux variant fortement selon les catégories d'utilisateurs, indique Emmanuel Brunet. Selon une étude du cabinet britannique e-Marketer, « plus d'un cinquième » des internautes français utiliseront cette année un bloqueur sur un terminal au moins: ordinateur, tablette ou téléphone mobile. Les tranches d'âges les plus touchées sont les 18-24 ans (près de la moitié utiliseront un bloqueur) et les 25-34 ans (38,3%).

La progression des bloqueurs a ralenti depuis 2016, moment charnière où les éditeurs de site se sont mis à refuser les internautes dotés de bloqueurs de pub sur leurs sites. Mais elle reste quand même élevée, selon e-Marketer : en 2018, le nombre d'internautes utilisant un bloqueur sur un terminal devrait encore augmenter de 9%, souligne l'étude. Face à cette situation, certaines sociétés proposent aux éditeurs de sites internet des solutions pour leur permettre de contourner les bloqueurs de pub, et faire parvenir quand même les messages publicitaires aux internautes.

Nouveaux outils anti-blocage

Inside Secure, une entreprise française qui s'est fait un nom notamment dans la gestion numérique des droits pour de grands noms comme HBO, Sky ou SFR (groupe Altice), va présenter la semaine prochaine de nouveaux outils antiblocage. Ces outils permettent de « brouiller » les messages envoyés par un site internet à l'ordinateur de l'internaute, pour que le bloqueur de pub ne parvienne pas à distinguer le contenu publicitaire et à le supprimer, explique à l'AFP Cyrille Ngalle, vice-président chargé de la protection des contenus chez Inside Secure. « Nous travaillons à faire en sorte que cette industrie (la publicité) survive, parce que sinon le modèle de distribution gratuite de contenus va mourir *», souligne-t-il.*

Chez Eulerian Technologies toutefois, Emmanuel Brunet croit plutôt à une meilleure collaboration entre annonceurs et éditeurs de sites pour éviter la fuite des internautes. Pour une marque, « il vaut mieux avoir une logique de partenariat » avec des éditeurs choisis plutôt « qu'une politique d'achat au kilo de bandeaux publicitaires », estime-t-il. « Il faut mieux intégrer la publicité dans les sites pour échapper aux bloqueurs et rendre les publicités plus acceptables », estime-t-il.

Un diagnostic qui est somme toute également celui d'Adblock Plus, l'un des bloqueurs les plus répandus, qui propose aux internautes une formule laissant passer des publicités dites « acceptables ». Selon Adblock Plus, seulement 25% de ses utilisateurs sont strictement opposés à toutes les annonces publicitaires, le reste, 75%, acceptant des publicités « pour contribuer à soutenir les sites internet *». Pour être inscrit sur la «* liste blanche *» des publicités acceptables, il faut en faire la demande à Adblock Plus, respecter certains critères techniques et … pour les plus gros annonceurs, payer Adblock Plus. Ces frais de licence constituent désormais la principale source de revenus d'Adblock Plus, explique l'entreprise sur son site internet*. »

ANNEXE 3 – Bibliographie

IAPP Publication – Privacy Program Management second édition

Informatique Technique – Hadoop Devenez opérationnel dans le monde du Big Data – Juvénal Chokogoue

Dunod – BIG DATA et Machine Learning – Les concepts et les outils de la data science - Pirmin Lemberger, Marc Batty, Médéric Morel, Jean Luc Raffaëlli

Dalloz – CYBERDROIT : Le droit à l'épreuve de l'internet – Christiane Féral-Schuhl

IAPP Publication – Foundation of Information Privacy and Data Protection : A survey of global concepts, laws and practices – Peter P. Swire, Kenesa Ahmad.

IAPP Publication – Gestion du programme de protection des données personnelles : Outils pour la gestion de la protection des données personnelles au sein de votre organisation

Le passeur – BIG DATA : penser l'homme et le monde autrement – Gilles Babinet

https://www.silicon.fr/big-data-igrandir-ou-mourir-169771.html

https://www.silicon.fr/un-cours-sur-le-bullshit-a-lere-du-big-data-169009.html

https://www.bigdataparis.com/documents/2020/BDG19_BD_19206.pdf

https://documents.epfl.ch/users/a/ay/ayday/www/mini_project/Sticky%20Policies.pdf

https://www.federation-eben.com/wp-content/uploads/2017/05/zdnet.fr_14032017_Big-Data_mais-Big-menaces-%C3%A9galement.pdf

https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html

https://www.dpms.eu/rgpd/associer-big-data-rgpd/

https://www.linformaticien.com/dossiers/comment-concilier-big-data-et-rgpd.aspx

http://www.mc2i.fr/A-grand-pouvoir-grande-responsabilite-le-Big-Data-vs-la-protection-des-donnees

https://www.daf-mag.fr/Thematique/reglementation-1243/Breves/Tribune-Quelle-place-big-data-ere-RGPD-328414.htm

https://www.cnil.fr/fr/definition/big-data

https://blog.httpcs.com/big-data-rgpd/

https://www.lebigdata.fr/fuites-de-donnees-augmentation

https://books.google.fr/books?id=KyRADwAAQBAJ&pg=PT115&lpg=PT115&dq=G29+%E2%80%93+WP136+%E2%80%93+Avis+4/2007&source=bl&ots=El3UyRF-G7&sig=ACfU3U3Wl8nIT--7Fd3me0YG7NNpnYWWYQ&hl=fr&sa=X&ved=2ahUKEwjF2dHOtNjnAhXL0eAKHdZfDDMQ6AEwA3oECAgQAQ#v=onepage&q=G29%20%E2%80%93%20WP136%20%E2%80%93%20Avis%204%2F2007&f=false


Patricia CHEMALI
Consultant indépendant
Data protection
Patricia.chemali[at]eDataPrivacy.fr
Tél. +33 6 28 71 63 59