BIG DATA

Bulletin de Veille Technologique (BVT) - Avril 2015

ARTCI
AUTORITE DE REGULATION DES TELECOMMUNICATIONS/TIC DE COTE D'IVOIRE

Sommaire

Glossaire

Agrégation de données: Regrouper des données à l'aide d'indicateurs comme la somme, la moyenne, etc.

Analytique L'analytique est l'application de l'informatique, de la recherche opérationnelle et de la statistique à la résolution de problèmes d'entreprise et industriels. Elle se réalise dans un système d'information. Discipline récente fortement liée à l'expansion de l'informatique, l'analytique s'est au début intéressée à l'usage des ordinateurs puis a évolué vers l'analyse des données. Les mathématiques sont essentielles aux algorithmes utilisés en analytique, qui tente d'extraire des informations utiles de grandes quantités de données

Anonymisation de données: (A fortiori personnelles) consiste à modifier le contenu ou la structure de ces données afin de rendre très difficile ou impossible la « ré-identification » des personnes (physiques ou morales) ou des entités concernées. Les enjeux sont éthiques, juridiques et associés à la bonne gouvernance.

Cloud Computing Le cloud computing, abrégé en cloud (« le Nuage »), ou l'informatique en nuage est l'exploitation de la puissance de calcul ou de stockage de serveurs informatiques distants par l'intermédiaire d'un réseau, généralement Internet. Ces serveurs sont loués à la demande, le plus souvent par heure ou par minute selon des critères techniques (puissance, bande passante, etc.) mais également au forfait. Le cloud computing se caractérise par sa grande souplesse : selon le niveau de compétence de l'utilisateur client, il est possible de gérer soi-même son serveur ou de se contenter d'utiliser des applicatifs distants

Cookie: Le cookie ou témoin de connexion est l'équivalent d'un petit fichier texte stocké sur le terminal de l'internaute. Existants depuis plus de 20 ans, ils permettent aux développeurs de sites internet de conserver des données utilisateur afin de faciliter leur navigation et de permettre certaines fonctionnalités. Les cookies ont toujours été plus ou moins controversés car contenant des informations personnelles résiduelles pouvant potentiellement être exploitées par des tiers

Data Scientist: Spécialiste de la data science. C'est la personne à qui une entreprise va confier ses données, variées et volumineuses (big data), et qui devra en tirer de la valeur, en ayant en tête les impératifs côté IT et côté business. Ce profil, encore en pleine définition, est né le jour où les organisations ont pris conscience qu'elles ne devaient plus simplement stocker leurs données pour des raisons légales, ayant tout intérêt à les exploiter de façon intelligente, et en quasi-temps réel. Ce métier requiert une connaissance en mathématiques, en base de données, en informatique et un pourvoyeur d'emplois.

Données non structurée: Ce sont des données présentes sous toutes les formes (images, textes, sons, vidéos). Elles nécessitent une structure de stockage spéciale longtemps sujet de recherche, d'où le succès du Big Data

Données structurée: Les données structurées sont des informations organisées et classées en vue de faciliter leur lecture et leur traitement. Elles sont sous forme de tableau. Un tableau Excel est une donnée structurée.

Framework: En programmation informatique, un framework est un ensemble cohérent de composants logiciels structurels, qui sert à créer les fondations ainsi que les grandes lignes de tout ou d'une partie d'un logiciel.

Science de données (Data Science) : C'est l'extraction de connaissance de données. Elle emploie des techniques et des théories tirées de beaucoup de domaines dont mathématiques, la statistique, la théorie de l'information et la technologie de l'information, le traitement de signal, des modèles de probabilité, l'apprentissage automatique, etc.

Snowden (Révélations d'Edward.) : Diffusion d'un important volume de données par un ex agent de la CIA, Edward Snowden. Ces données concernent la surveillance et l'espionnage d'internet et des téléphones mobiles par l'agence américaine de renseignement, la NSA (National Security Agency).

<u>Rappel - Octet</u>

1 **kilooctet** (ko)= 10^3 octets= 1 000 octets

1 **mégaoctet** (Mo)= 10^6 octets= 1 000 ko= 1 000 000 octets

1 **gigaoctet** (Go)= 10^9 octets= 1 000 Mo= 1 000 000 000 octets

1 **téraoctet** (To)= 10^{12} octets= 1 000 Go= 1 000 000 000 000 octets

1 **pétaoctet** (Po)= 10^{15} octets= 1 000 To= 1 000 000 000 000 000 octets

1 **exaoctet** (Eo)= 10^{18} octets= 1 000 Po= 1 000 000 000 000 000 000 octets

1 **zettaoctet** (Zo)= 10^{21} octets= 1 000 Eo= 1 000 000 000 000 000 000 000 octets

1 **yottaoctet** (Yo)= 10^{24} octets= 1 000 Zo= 1 000 000 000 000 000 000 000 000 octets

Introduction

« Target savait à quel stade de sa grossesse se trouvait la jeune . Avant même que son père ne s'en rende compte! »

L'histoire se déroule dans un supermarché des Etats-Unis. Un homme en colère demande à voir incessamment le directeur : « Ma fille, encore au lycée reçoit des promotions sur des vêtements de bébé et des berceaux, voulez-vous qu'elle tombe enceinte ? » . Le directeur ébahi présente alors ses excuses sans vraisemblablement comprendre ce qui se passe. Effectivement, le magasin envoyait des publicités et proposait des réductions sur des berceaux, des grenouillères, etc. Quelques jours, plus tard, le directeur du magasin rappelle encore pour s'excuser et à son double étonnement tombe sur un père qui culpabilise en lui disant : « Je vous dois des excuses, je n'étais pas au courant de tout ce qui se passe chez moi, ...ma fille va accoucher en août ». Cette histoire rapportée en 2012 par le New York Times a fait le tour du web.

Comment le magasin savait à quel stade de sa grossesse se trouvait la jeune ; avant même que son père ne s'en rende compte?

Pour cela, Target, la chaîne de magasins en question a listé 25 produits susceptibles d'être achetés par les femmes enceintes avec des quantités différentes de la normale.

Ces prédictions sont établies à partir de données de plus en plus volumineuses provenant de différentes sources (données de navigation internet, capteurs, mobiles, achats, etc.) et de puissants algorithmes. L'ensemble constitue le « Big Data ».

Au départ, l'apanage des géants du web, l'usage du « Big data » va se démocratiser pour se répandre dans tous les secteurs.

Les pays émergents, en l'occurrence ceux de l'Afrique ne restent pas en marge. Les données issues des téléphones mobiles deviennent une mine d'or pour les grands groupes, laboratoires et consommateurs de données.

Cependant il va falloir rester vigilant. L'utilisation du Big Data reste mitigée dans la mesure où il tend à s'immiscer dans la vie privée des individus ,de façon difficilement contrôlable. Aussi, les assertions qui découlent de l'analyse dans le Big Data ne sont pas vérifiées dans l'absolu.

Il s'agira dans ce bulletin de vous entretenir en long et en large sur le Big Data, ses applications, ses défis et notre position d'arbitre en tant que régulateur.

Contexte

Selon IBM, nous générons chaque jour « 2,5 trillions d'octets de données. Et 90% de ces données ont été créées au cours des deux dernières années seulement. Ces données proviennent de partout notamment de capteurs utilisés pour collecter les informations climatiques, de messages sur les sites de médias sociaux, d'images numériques et de vidéos publiées en ligne, d'enregistrements transactionnels d'achats en ligne et de signaux GPS, de téléphones mobiles, etc. » Des entreprises vont faire de ces données leur fer de lance à travers des analyses complexes pour conquérir le marché. Le terme général de Big Data, en français « mégadonnées » prend son envol.

Comment cela est-ce possible ? A cause de deux révolutions informatiques.

La première concerne les capacités de stockage. 1 Mégaoctet coutait en moyenne 3500$ en 1980. Aujourd'hui on peut acheter pour 20$ des capacités 8000 fois plus grandes. Les capacités de stockage ne sont plus une contrainte compte tenu de leur coût insignifiant.

La seconde innovation ayant boosté le Big Data : L'informatique en parallèle et distribué. Plutôt qu'un seul superordinateur résolve un problème, il est possible de diviser ce problème en plusieurs sous parties gérées chacune par des milliers d'ordinateurs standards.

Cette performance associée à la baisse des coûts de la puissance de calcul, du transfert de données et à des débits internet de plus en plus rapide a engendré l'informatique en nuage ou « cloud computing ».

A cela, on pourra ajouter l'émergence de la téléphonie mobile et de l'internet. L'UIT dans son rapport statistique de 2014 publie que le monde compte, fin 2014, près de 7 milliards d'abonnements au cellulaire mobile et 3 milliards d'internautes.

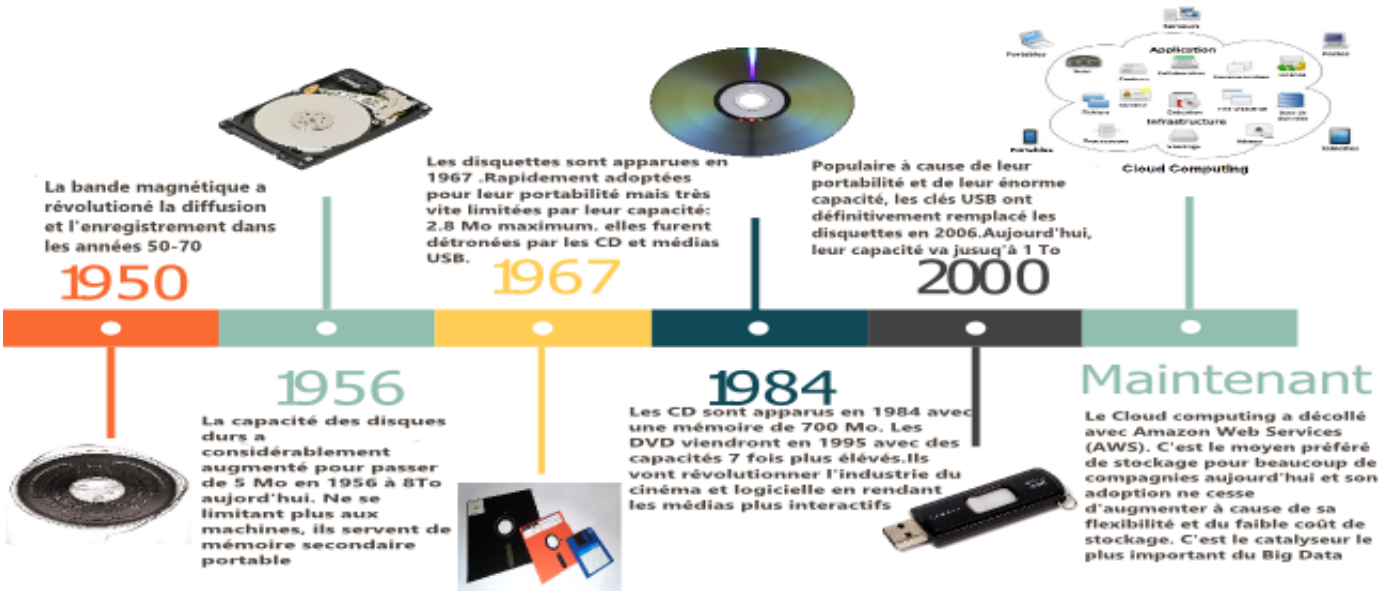
Cette affluence de données va demander des moyens de gestion, de stockage et d'analyse différents de ceux utilisés par les bases de données traditionnelles.

Les systèmes d'informations décisionnelles (Business Intelligence) essaieront de consolider toute cette information et la présenter dans un format convenable pour le décideur. On parle de cube et d'entrepôts de données. Mais tous ces systèmes restent insuffisant devant les données de types textes (commentaires, blog, articles de presse), images et vidéos qui regorgent d'informations pertinentes difficiles à extraire et les plus nombreuses à cause de l'explosion des réseaux sociaux. Ces données sont dites non structurées. Aussi le besoin de stocker les données en heure est de plus en plus récurrent. Les systèmes BI sont limités à cause des performances de stockage que cela requiert.

« 7 milliards d'abonnements au cellulaire en 2014. 3 milliards d'internautes», (UIT)

Histoire du stockage de données

(WIKIPEDIA)



Définition

« 8 zettaoctets de données générées dans le monde d'ici fin 2015 » (IDC, EMC, Gartner)

Le Big Data se définit comme un ensemble de données dont le grand volume nécessite des moyens et outils d'analyse adaptés différents de ceux utilisés par les bases de données classiques et systèmes d'informations.

La définition du Big Data prend tout son sens autour de quatre axes d'analyse. En effet, le Big Data c'est la capacité de traiter des données diverses (variété) tels que l'historique de navigation, les déplacements dans une ville ou un magasin pour proposer en temps réel ou presque (vélocité) des offres personnalisées indispensables (véracité) pour des milliers de clients (volume).

Volume

Le volume peut être défini comme l'élément le plus capital dans le Big Data. Il est plus concluant par exemple d'évaluer l'efficacité d'un traitement médical sur des données provenant d'une large population que sur quelques patients.

Ainsi, selon IBM, Twitter par exemple utilisera ses 12 Téraoctets de tweets quotidiens pour une analyse sur des produits ou tendances mondiales. Selon une étude IDC de 2011, sponsorisée par EMC Gartner, les données numériques ,créées dans le monde, estimées à 1,8 zettaoctets en 2011 devraient atteindre 8 zettaoctets en 2015.

Vélocité

Le traitement des données doit être effectué à mesure que les données sont collectées pour favoriser la vitesse de la prise de décision.

La vélocité est définie comme la fréquence à laquelle les données sont générées, capturées et partagées. Les technologies émergentes comme le cloud et les algorithmes de type « analyse de flux de données » sont capables de répondre à de telles exigences. Ce qui augmente la souplesse avec laquelle les organisations peuvent répondre aux changements du marché, aux préférences clients ou cas de fraude. La course à la vitesse est un facteur clé afin d'obtenir un avantage concurrentiel dans certains secteurs comme la bourse

Variété

Le Big Data inclut tout type de données. Textes, données de capteur, déplacements, données mobiles issues de station BTS, enregistrement d'appels, enregistrements audio et vidéo, fichiers de logs, flux de clics sur les sites Internet, etc.

Ces données non structurées différentes des données structurées (sous forme de table, issues des bases de données traditionnelles) mettent les structures de stockage et les analystes devant un réel défi. Les modélisées dans une forme d'agencement adaptée pour traitement et analyse peut demander du temps et un effort considérable.

Ces données longtemps délaissées ont vu leur potentiel croître grâce aux nouvelles technologies induites par le Big Data qui facilitent leur analyse. Par exemple, nous pouvons citer les analyses d'opinions et de sentiment à partir des commentaires des réseaux sociaux qui permettent de s'enquérir des réputations et déceler des tendances sur des mots ou évènements.

Véracité

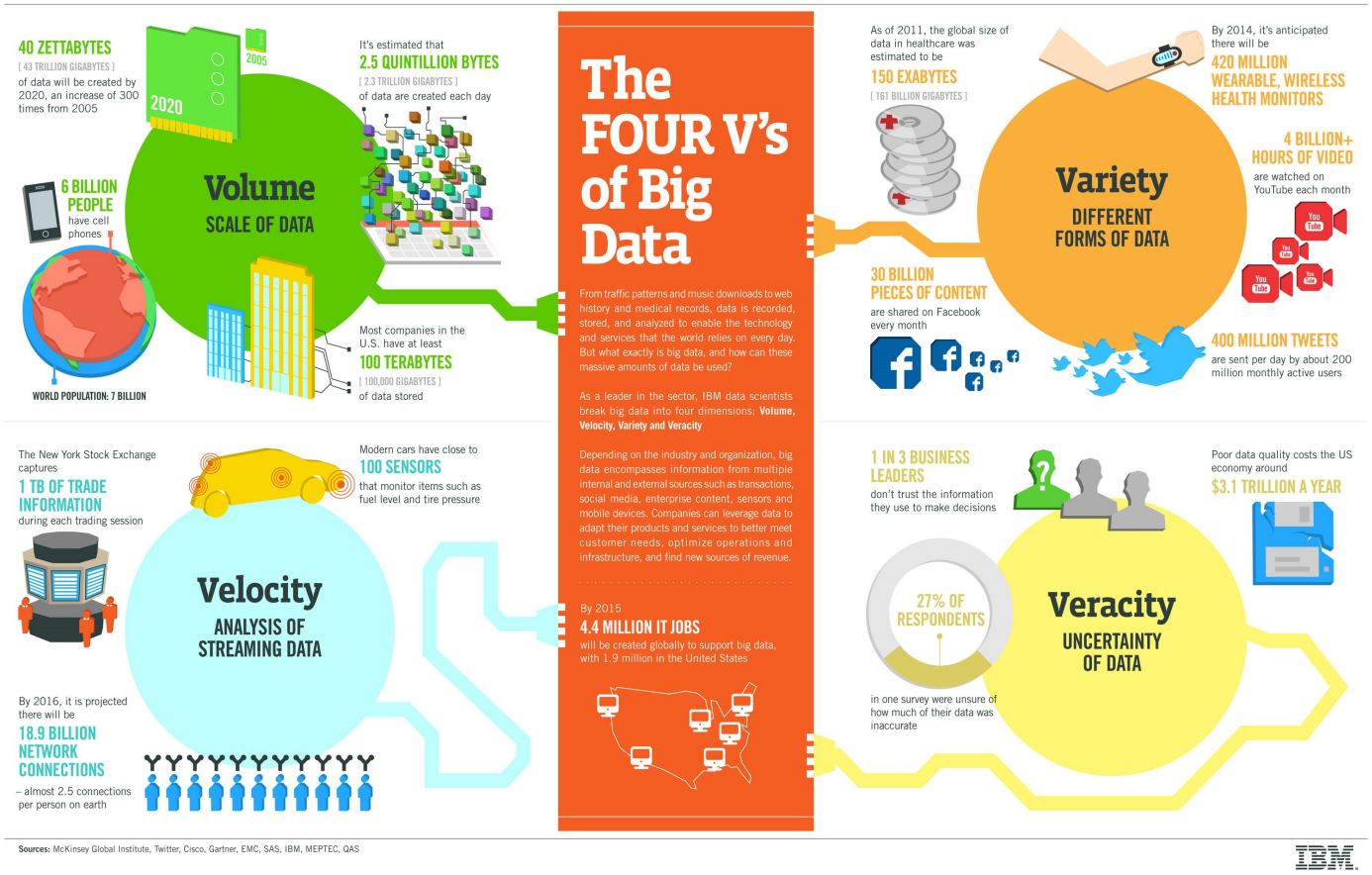
Quelle est la précision des données sur lesquelles nous voulons bâtir notre modèle de prise de décision ?

Influencé par les trois précédents V, le Big Data tend à contenir un certain niveau d'incertitude, d'incohérence, parfois d'ambiguïté et d'aberration. Le degré d'incertitude et la qualité des données peuvent varier mais ils doivent être pris en compte dans tout projet Big Data. Des tests statistiques de significativité à grandes échelles doivent être mis en place au risque de perdre toute crédibilité. Le système final doit être à même de distinguer, évaluer, peser la véracité et le niveau de significativité des données.

On parle aussi d'un cinquième V pour « *Valeur* ». En effet, il serait futile de réaliser des analyses complexes sans en tirer une information utile.

L'accent sur l'un ou l'autre des V est spécifique à chaque application.

Certaines applications vont se concentrer que sur une petite quantité de données, les analyser de façon continue mais avec des algorithmes complexes (proposition de produits basée sur les déplacements en temps réel, détection de fraude). Tandis que d'autres seront réalisés en période d'activité creuse à cause des gros volumes de données qu'elles mobilisent (systèmes de recommandations en ligne).



Les 4 Vs du Big Data (IBM)

Les sources du Big Data

Selon la vision de l'équipementier Ericsson, plus de 50 milliards d'objets seront connectés d'ici 2020, indépendamment de leur origine, favorisant ainsi l'Internet de toute chose, de tout lieu. Tous ces appareils vont mesurer, capter, générer et communiquer des données d'une certaine taille et structure (structurée ou non structurée).

L'UIT estimait à sept (7) milliards les souscriptions mobiles dans le monde à la fin de 2014, et chacune de ces souscriptions est en même temps créatrice et consommatrice de données. Aujourd'hui, plus de 3 milliards de personnes utilisent l'Internet et les souscriptions au large bande mobile, en particulier, ont été propulsées de 268 millions en 2007 à 2,3 milliards en 2014. Chacun de ces consommateurs contribue au déluge de données avec les SMS, les appels, les photos, les vidéos et les messages postés sur les sites de médias-sociaux, les emails, les recherches, les clics sur les liens et les publicités, les courses en ligne et les payements mobiles, ou la géolocalisation grâce aux traces laissées par les smartphones ou les connexions sur les bornes d'accès Wi-Fi.

Aussi, le développement du large bande favorise l'utilisation de services gourmands en bande passante (vidéo, vidéo téléphonie, etc.). Ce qui a pour conséquence, une explosion de données et du trafic sur les réseaux IP.

La philanthropie de la donnée et l'open data

L'ouverture des données (Open Data en anglais) est une philosophie visant à rendre des données numériques (du secteur publique ou privé) accessibles à tous et à s'affranchir des restrictions sur le droit d'accès et de réutilisation.

Globale Pulse, l'initiative « Big data pour le développement » de l'ONU, compte sur l'ouverture des données où les fournisseurs de services de télécommunications partage les données des consommateurs de façon sécurisée pour un accès ouvert à tous.

Plusieurs pays ont déjà adhéré à l'initiative de l'open data en fournissant des données sur la santé, le transport, l'agriculture, etc. Des utilisateurs vont en retour opérer des traitements et fournir des résultats via des graphiques interactifs qui ne seraient visibles et déterminées de façon brutes, sans manipulation.

Application

Beaucoup de secteurs sont ou pourraient être fortement impactés par le Big Data.

Les entreprises et l' analytique

Visa est soupçonné d'utiliser un modèle mathématique pour savoir si ses clients vont divorcer. En effet, elle s'appuie sur le fait que les personnes en instance de divorce ont tendance à payer leur facture en retard.

En 2012, Facebook enregistrait près de 2.5 milliards de « j'aime » par jour. Une étude publiée par l'académie nationale des sciences des Etats-Unis (PNAS) stipule qu'un algorithme est capable de déterminer la personnalité d'un individu à partir du nombre de « j'aime ». Selon la même étude, avec « 10 J'aime », l'algorithme vous connait mieux qu'un collègue, avec 500, mieux qu'un conjoint.

Ces exemples montrent à quel point les entreprises cherchent de plus en plus à anticiper ou à prévoir les événements à venir, que cela soit dans l'évolution des risques (risque sécuritaire par exemple lié aux cyberattaques), la lutte contre la fraude, la gestion des ressources humaines et l'analyse des comportements clients.

Le cabinet Deloitte argue sur la tendance analytique en disant qu'elle semble naturellement induite par un contexte de prolifération active des données et de digitalisation croissante (web, applications mobiles, réseaux sociaux, objets connectés). Dans un contexte économique tendu, cette tendance se justifie également par le besoin de disposer d'analyses de plus en plus fines et variées au sein des organisations. Impulsés par les courants du Big data et du Cloud, les usages analytiques ne cessent de se développer. Les entreprises de développement logiciel, de conseil informatique et les startups prolifèrent et accentuent de plus leur offre Big Data et « *analytique* » dans différents secteurs. Ce, pour une utilisation conviviale et une aide à la décision assez flexible et réactive.

La mobilité

La mobilité est un challenge des temps modernes. Les villes se développent de plus en plus et les professionnels du secteur du transport doivent s'adapter et innover afin d'accroître l'efficacité et la durabilité.

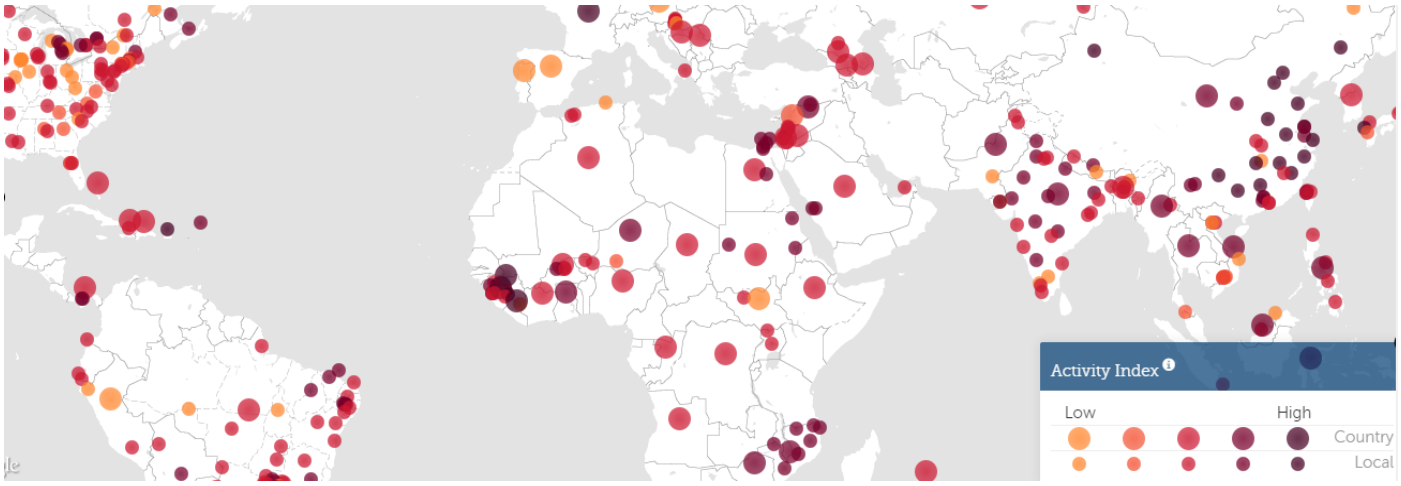
Les cartes de validation dans les transports , les données provenant des capteurs GPS, les données de mobiles représentent autant de vecteurs d'information. Le Big Data valorisera les informations générées afin de prédire en temps réel les embouteillages, les engorgements les fortes zones de trafic et mieux orienter les conducteurs.

Il est possible par exemple de reconstruire les modèles de mouvement d'une communauté à partir des données provenant des communications de téléphones mobiles. Cette information peut être utilisée pour visualiser des rythmes quotidiens de déplacements vers et à partir de la maison, du travail, de l'école, des marchés mais aussi de réaliser des applications liées à la propagation d'une maladie ou aux mouvements d'une population touchée par une catastrophe.

Genève le vendredi matin: 08h00 à gauche, 12:h00 à droite. (villevivante.ch)

Les mouvements des citadins sont représentées par la densité des lignes., obtenues grâce aux traces ... des téléphones mobiles. (villevivante.ch)

« Avec 10 ''J'aime'' sur Facebook, un algorithme est capable de mieux vous connaître qu'un collègue de bureau. Avec 500 ''J'aime'', mieux qu'un conjoint»



Alertes relatives à différents virus ou maladies dans le monde entier

HealthMap composée d'une équipe de chercheurs, d'épidémiologistes et développeurs traite différentes sources officieuses en ligne à l'aide d'algorithmes d'analyse de données. Les informations extraites permettent de surveiller et de détecter en temps réel les menaces de santé publique. (healthmap.org)

Des communautés de conducteurs partagent beaucoup d'information sur l'état de la route explicitement via des applications dédiées (ex : Waze) ou implicitement avec les traces laissées par leur téléphone mobile. Ce genre d'initiatives procure un énorme gain de temps dans les trajets.

La publicité en ligne

Le Big Data constitue aujourd'hui une mine d'or pour les publicitaires. En effet, les clients et potentiels prospects internet vont inconsciemment laisser une empreinte digitale sur Internet révélatrice de leur comportement d'achats et surtout des centres d'intérêts.

La plupart des internautes ont déjà eu cette sensation d'être traqué par les publicités. En effet, la génération de ces contenus est effectuée sur la base d'algorithmes scientifiques de recommandations et de re-ciblage. Ces algorithmes vont à partir de *cookies*, des statistiques de connexions, du temps passé sur les réseaux sociaux, des pages visités, des produits visualisés… en un mot de notre parcours sur Internet, identifier ou détecter les produits qui pourraient nous intéresser. Parfois même, ce sera pour nous relancer sur des produits que nous avons regardé sans pour autant acheter et proposer des offres promotionnelles très alléchantes spécifiques et uniques sur ces produits. Ainsi au lieu de bombarder la population avec des publicités génériques, les entreprises proposent des offres plus ciblées, donc plus efficaces.

En Côte d'Ivoire, Jumia, entreprise de E-commerce utilise les services proposés par Facebook et les réseaux publicitaires de Google et Apple pour accroitre sa présence digitale.

Ce qu'il faut retenir, c'est que nos traces de navigation sur Internet constituent aujourd'hui l'essentiel du Big Data qui l'utilise à des fins commerciales et marketing.

La santé

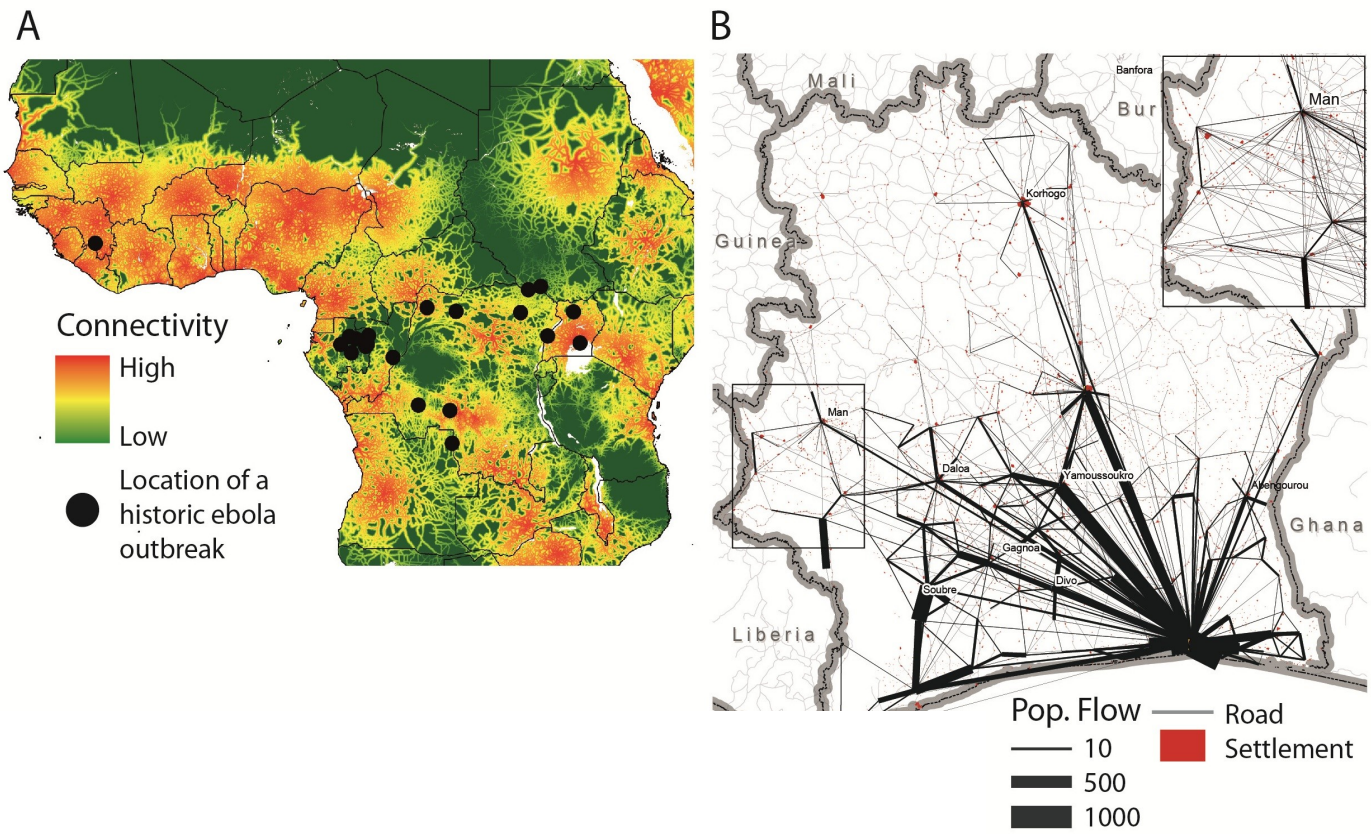
Dans le domaine de la santé, l'analyse des données est très importante dans la mesure où l'historique des données d'un patient fournit des éléments pertinents dans la prise de décision finale du personnel de santé sur le traitement.

Si les données dans le domaine médical étaient collectées assez régulièrement, (consultations, mammographies, radiographies, etc.), les besoins des populations pourraient être prédits et mieux circonscrits. Avec l'Internet des objets et l'informatisation des services administratifs, le suivi électronique des patients commence à se généraliser. Le taux de glucose des diabétiques peut être suivi avec des glucomètres connectés. Les appareils connectés comme les traqueurs d'activité transmettent des informations cardiaques ou générales. Des séries temporelles sur l'état de santé des patients sont à chaque fois envoyées à un centre de traitement. Ce qui permet de détecter des anomalies et prédire l'état futur du patient.

Le Big Data intervient dans le stockage et le traitement des informations transmises. Le Big Data, grâce à ses performances, peut détecter des corrélations ou patterns entre les signes cliniques d'une population, les déclenchements et la propagation d'un virus. Ces innovations pourraient réduire de façon significative les investissements réalisés dans le domaine de la santé.

Un exemple concret serait de déterminer la propagation du Virus Ebola à partir du croisement entre les données mobiles issues des stations de bases et des mobiles, aussi des données des aéroports et des gares de transports avec des données et indicateurs relatifs à la santé. A l'aide de cet ensemble, les flux migratoires les plus risqués ainsi que les hot-spots de la maladie pourraient être identifiés.

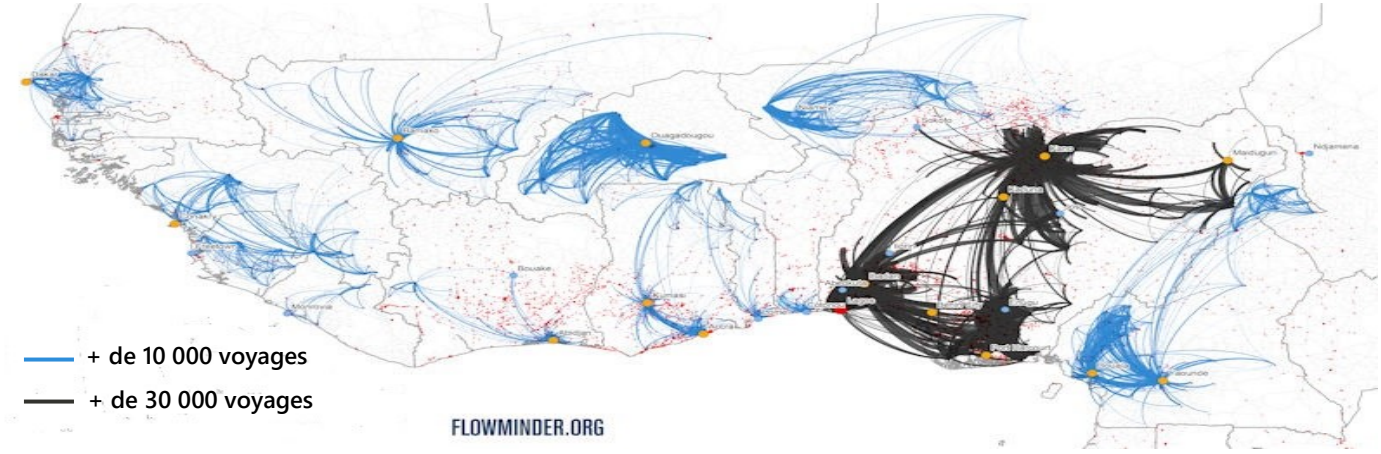
Big Data en Afrique



A) Carte montrant les foyers d'Ebola depuis 1976 (points noirs) superposé avec une carte de connectivité mesuré à partir du temps de déplacement à la localité . Plus la connectivité est élevé, moins est le temps de voyage. Préalablement, la fièvre Ebola apparaissait dans les zones à faible connectivité . L'apparition de la fièvre en Guinée provoque une cassure car c'est une zone à forte connectivité et densément peuplée.

B) Flux de 500 000 utilisateurs de téléphonie mobile en Côte d'Ivoire. La région de man enregistre plusieurs mouvements sur une distance d'au moins 20km en moyenne avec au moins 10 voyageurs par trajet

(http://currents.plos.org/outbreaks/article/containing-the-ebola-outbreak-the-potential-and-challenge-of-mobile-network-data/)



Estimation de flux de voyages dans la sous-région à partir des données de téléphone mobile.

La Côte d'Ivoire , le Sénégal et le Kenya présente les mêmes caractéristiques. Les lignes en bleu représentent 10 000 voyages sur au moins 20 km. Au Nigéria, les lignes noirs représentent plus de 30 000 voyages estimés entre deux localités. (http://www.worldpop.org.uk/ebola/)

L'Afrique et les pays en développement sont aussi très impliqués dans le phénomène Big Data.

Cette implication est favorisée par la percée de la téléphonie mobile dans ces régions du monde. Selon une étude menée conjointement par GSMA et Deloitte, le taux de pénétration du mobile serait entre 70 et 80% en Afrique subsaharienne en 2015. Il est de 97.05% en Côte d'Ivoire en fin 2014.

Le mobile est aujourd'hui utilisé en Afrique subsaharienne pour communiquer, transférer de l'argent, acheter et payer les factures.

Chaque communication avec un téléphone mobile génère des traces appelées Call Detail Record (CDR). Les CDRs contiennent la localisation des stations relais d'émission et de réception de signal, l'heure de la communication et sa durée. La localisation des stations relais renseigne approximativement sur celle de l'appelant ou l'appelé car elles sont généralement situées à 1-2km de l'usager en ville et 3 – 7 km en zone rurale.

Le Big Data est amène de transformer ces données en informations utiles pour améliorer le cadre de vie des Africains et des pays en développement dans différents secteurs :

Les temps de communications peuvent être utilisés pour estimer le niveau de vie et de dépense des ménages. Le niveau de vie est supposé aisé quand le temps de communication est élevé. Ce suivi sur les temps de communication mobiles peut être révélateur de tendances . Des changements soudains dans la façon de communiquer des consommateurs peuvent permette de détecter assez rapidement des problèmes macro comme des crises ou bien même la perte de consommateurs

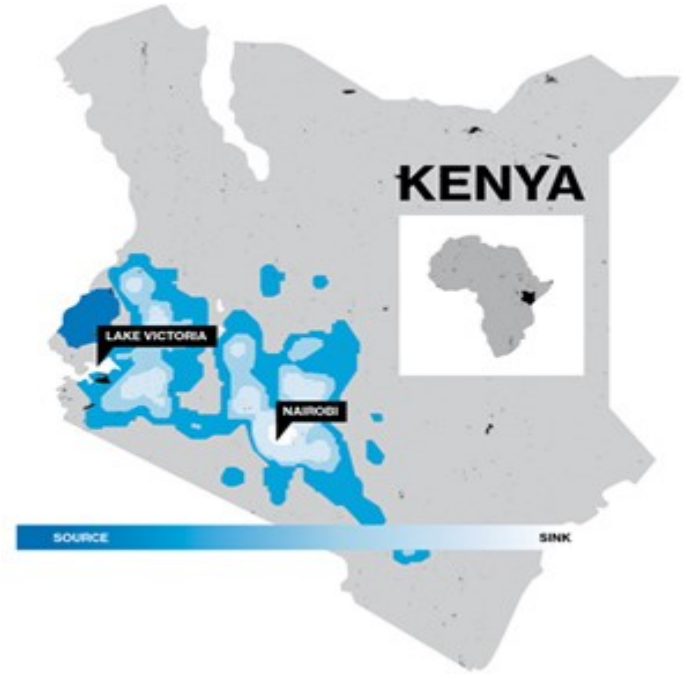
Citons quelques exemples :

Socio-économique

Dans une autre étude, les chercheurs ont utilisé des CDR anonymisées de cinq (5) millions de clients de l'opérateur de téléphonie mobile Orange CI, entre Décembre 2011 et Avril 2012, pour cartographier les niveaux de pauvreté en Côte d'Ivoire ainsi que le niveau d'activité des abonnés mobiles. Grâce à ces données, les niveaux de pauvreté de onze (11 régions) de la Côte d'Ivoire ont été quantifiés. Cette estimation a été validée par rapport à un indice de pauvreté multidimensionnelle créé par l'Université d'Oxford. L'application des technologies du Big Data aux CDR peut constituer un complément d'information par rapport aux enquêtes nationales d'estimation de l'économie et de la croissance.

Santé

En 2012, une étude menée par des chercheurs sur près de 15 millions d'abonnés de téléphonie mobile au Kenya a révélé les foyers d'origine et les possibles endroits de propagation de la malaria dans le pays. Les foyers d'origine infectieux étaient situés vers le Lac Victoria et les voyageurs transportaient la maladie pour la propager dans de potentiels foyers comme la capitale Nairobi.



Carte du Kenya résultant de l'analyse de données de téléphone mobile .

Montrant les principaux foyers d'origine de la Malaria (vers le lac Victoria en bleu sombre) et les potentiels lieux de propagation (la capitale Nairobi en bleu clair) (. http://www.technologyreview.com/featuredstory/513721/big-data-from-cheap-phones/)

Mobilité

A cause de l'urbanisation galopante des pays en développement, les routes et systèmes de transport deviennent de plus en plus saturés. Le challenge D4D organisé par Orange a bien voulu adressé cette problématique en fournissant les CDRs à des chercheurs du monde entier. Un laboratoire du Géant IBM a essayé d'analyser ces questions avec les CDRs fournis et grâce à des techniques d'analyse très poussées. Leurs résultats ont permis de proposer une conclusion partielle à la congestion de la ville. Les chercheurs proposent l'ajout de quatre (4) voies à l'infrastructure existante et l'extension d'une autre voie. Ces aménagements pourraient réduire le temps de voyage de 10%

Défis et enjeux

du BIG DATA

« LE Big Data peut induire des coûts importants ….Le cloud Computing s'impose comme la solution. »

Technologie et exigences

Infrastructure

L'infrastructure informatique des entreprises doit accorder la priorité aux informations qui circulent en temps réel mais aussi sur les complexes requêtes d'alimentation de bases de données s'effectuant généralement nuitamment.

L'infrastructure doit être optimisée spécifiquement pour les besoins en analytique – avec des processus métiers automatisés – afin de répondre aux demandes croissantes du Big Data.

Les géants du web utilisent parfois des logiciels et infrastructures propriétaires. Yahoo, Google ont bien voulu livrer à la communauté leur expertise et outils en matière de gestion de données. Ce partage fut considérer comme une révélation et ces outils servent aujourd'hui en quelque sorte de standard dans le domaine du Big Data.

Le plus populaire est le framework Hadoop. Ce framework est destiné à faciliter la création d'applications distribuées et échelonnables. Dans la même lignée, nous pouvons citer d'autres applications comme Storm, Spark, etc.

Hadoop offre un système de fichiers distribué dénommé HDFS qui permet de stocker de façon efficiente sur des milliers de nœuds de calcul. Ce système distribué a pour avantage de réaffecter automatiquement les routines et tâches quotidiennes vers d'autres nœuds en cas de défaillance de certains nœuds. Ce qui offre une simplicité de conception, d'évolutivité et surtout des performances en termes de latence et de débit.

Pour le traitement des données stockées, comme signifié plus haut, le Big Data offre des solutions pour s'affranchir des contraintes imposées par les systèmes de gestion de base de données traditionnelle. Map reduce est un modèle de programmation combiné généralement avec Hadoop et inventé par Google permettant de manipuler de grandes quantités de données (supérieure à 1 To) en les distribuant dans un ensemble de machine. Il répartit ses opérations sur chaque nœud augmentant ainsi la rapidité. Il trouve un franc succès chez les grandes compagnies comme Facebook ou Amazon. Il représente une solution incontournable pour le traitement des données non structurées comme le texte, les images, etc.

Cloud et Big data

Le Big Data peut induire des coûts importants lorsqu'il est déployé dans notre propre infrastructure. L'ingénierie en plus de se concentrer sur le cœur de métier devra aussi gérer le stress relatif à la maintenance de l'applicatif.

En effet, se doter de serveurs qui ne seront qu'utilisés parfois à moins de 5% de leur capacité ou au pire insuffisant pour le développement de certains projets peut s'avérer très couteux, difficile à maintenir et peu extensible. En plus des besoins gourmands du Big Data, il y a l'expertise associée qu'il faudra développer de l'installation à la maintenance du système analytique.

Raison pour laquelle le Cloud computing s'impose comme la solution. L'élasticité du nuage offre des services évolutifs, supporte des montées en charge et permet le dimensionnement automatique des applications. L'allocation de ressources se fait à la demande et évolue en même temps que les besoins. Les coûts d'entrée et de gestion restent mieux maitriser. En plus les responsabilités restent bien définies car des contrats sont établis comblant les satisfactions de chacune des parties.

C'est pourquoi les leaders des services de cloud computing (Azure, AWS, Google cloud) offre déjà des suites et services dédiés spécifiquement à la manipulation des données à grande volumétrie ou ayant des exigences de traitement en temps réel.

Le cloud garantit une énorme flexibilité et un énorme gain de temps accompagné d'une maitrise et clarté de budget.

Ainsi, le métier reste focalisé sur les attentes client ou de l'organisation tandis que l'ingénierie continue de développer des services liés au cœur de métier en dépensant très peu d'énergie sur l'infrastructure. Le Big Data déployé grâce aux technologies du cloud offre des capacités analytiques très poussées permettant de guider les décideurs à tous les niveaux de l'organisation.

Challenges

Vie privée

La vie privée, définie comme le droit des individus à contrôler la divulgation des informations les concernant est un pilier de la démocratie .Des mécanismes de protection de données doivent être mis en place pour éviter de compromettre ce droit durement éprouvé au regard des révélations de Snowden qui ont montré le rôle des technologies liées au Big Data dans l'intrusion de la vie privée.

Les individus peuvent eux même ne pas être conscients du type d'informations qu'ils génèrent vu les nombreuses interactions sur le mobile et les réseaux sociaux. Il est donc important de définir un cadre qui circonscrirait l'acquisition, la collecte et l'exploitation de ces données.

L'anonymisation et l'agrégation des données sont très utilisées par les pourvoyeurs de données pour parer aux problèmes de vie privée. Cependant ces deux recommandations revêtent des défis à relever

Pour l'anonymisation, il faudra garantir que l'information anonymisée est irréversible (impossibilité de retrouver l'information initiale).

Concernant l'agrégation de données, il s'agit de compromis. Il faut s'assurer que les agrégats résultants sont pertinents et contiennent l'essentiel de l'information (avec un risque de perte) voulue étant entendu que si les données étaient transmises à un niveau plus détaillé, des analyses plus complètes seraient possibles mais le risque lié à l'atteinte de la vie privée serait élevé.

Fracture Numérique

Dans les pays avec un fort taux de pénétration du mobile et une bonne présence d'internet, les citoyens produiront plus de données tandis que dans les autres, les données vont provenir d'initiatives ou programmes et nationaux et des instituts.

Une attention particulière doit être portée vers les pays qui produisent moins de données et n'ont pas de compétences et de capacités en termes d'analyse de données ; afin d'éviter de creuser une fois de plus le fossé numérique existant.

Le projet « GlobalPulse »des Nations Unies est une solution à ce problème.

Analytique ,compétence et gouvernance de données.

Le processus d'analyse de données dans le Big Data afin d'extraire de l'information pertinente comporte des risques qui peuvent jouer sur la qualité et la précision des résultats. L'analyse des données du Big Data pose dans un premier temps des défis liés à la méthodologie et à l'interprétation des résultats. Ensuite, Le métier de data scientist , n'est pas encore vulgarisé , l'offre est encore très inférieure vis-à-vis de la demande.

Les entreprises cherchent à mieux cerner leur patrimoine de données, dans un contexte toujours aussi grandissant en termes de volumétrie et des dépenses liées. On parle de gouvernance de données. Pour cela, le rôle du « Chief Data Officer » se ressent de plus en plus dans les organisations afin de créer un cadre de partage de données, d'appropriation par les différentes directions. Le Chief Data Officer aura également pour rôle de définir les règles et usages liés à l'exploitation des données ainsi qu'à la protection de ce patrimoine

« L'anonymisation et l'agrégation des données sont très utilisées par les pourvoyeurs de données pour parer aux problèmes de vie privée. »

« le rôle du ''Chief Data Officer '' se ressent de plus en plus dans les organisations »

Rôle du

Régulateur

Il est clair que le Big Data, compte tenu de ses qualités si puissantes façonnera tous les secteurs. Et vu qu'on ne peut occulter les abus (ex: atteinte à la vie privée) que peut engendrer son application, la régulation s'impose. Ici, nous essaierons de définir quatre axes sur lesquels le régulateur devra se pencher :

La protection de la vie privée et la sensibilisation des utilisateurs du Big Data.

Des pays, conscient de l'usage parfois problématique des données de leurs citoyens vont voter des lois sur la protection des données à caractère personnel. Tel est le cas de la Côte d'Ivoire au travers de la Loi n°2013-450 du 19 juin 2013 relative à la protection des données à caractère personnel.

Cette loi pourrait être renforcé par un dispositif de sensibilisation et d'information. Il s'agira d'informer les utilisateurs de la données, des dangers du Big data, des conséquences et aussi des garanties à mettre en place dans leur système ou son application avant de commencer à utiliser les données.

Ainsi les utilisateurs du Big Data pourront faire évaluer leurs applications et montrer qu'elles offrent des garanties en matière de protection de données. Requérir le consentement pour l'exploitation des données des personnes concernées sera évident.

Vu que les populations portent rarement plainte sur l'utilisation abusive de leurs données (par ignorance), l'Etat et les citoyens pourront compter sur le régulateur qui doit mettre tout en œuvre sur la sensibilisation.

Les prédictions probabilistes et la violation des droits de l'Homme

La prédiction constitue le nerf du Big Data. Il est très important de la contrôler afin d'éviter des dérapages.

Prenons l'exemple d'un gouvernement qui décide de surveiller un individu, non sur la base de ses interactions antérieures avec les services de police mais plutôt sur la base de prédictions faites à partir du Big Data. Cela constitue une entrave au principe de liberté.

Le régulateur, arbitre, doit montrer la ligne rouge aux gouvernements afin qu'ils ne se basent sur les prédictions pour punir ou surveiller.

Contrôler l'expertise du Big Data

Le Big Data, incontournable, fait appel à une expertise de la part de grands consommateurs de données mais aussi de la part de simple citoyen qui ont par exemple été victimes de fausses prédictions à leur rencontre et qui ont besoin d'experts dotés de compétence en analyse de données pour se défendre.

Le régulateur pourra encadrer des experts qui devront être formés sur les techniques liées au Big Data, les algorithmes mais aussi aux dispositions éthiques et légales . Ils devront faire vœux d'impartialité de confidentialité et de professionnalisme.

Ils pourront être référencés, sensibilisées sur la responsabilité d'utiliser les données . Ils seront aussi aptes à implémenter des systèmes de collecte, d'analyse et de restitution au niveau des entreprises consommatrices de données tout en respectant la législation en vigueur.

L'économie de données.

La quantité volumineuse des données est détenue par très peu d'entités. Le régulateur se doit de prendre des dispositions afin d'éviter que ce secteur se transforme en monopôle de vente de données. Le régulateur doit permettre l'accès aux banques de données à des prix justes et raisonnables. Des lois doivent être établies pour prévoir une exclusion juridique de droit en cas d'abus. Il faudrait que l'économie qui se crée autour des données contribue à l'amélioration de la qualité de service, engendre de nouveaux modèles collaboratifs et incite à l'innovation

« Informer les utilisateurs de la donnée sur les dangers du Big Data, les conséquences et les garanties à mettre en place. »

Conclusion

La majorité des actions de notre vie quotidienne laisse des traces numériques. Retirer de l'argent, naviguer sur Internet, « Like » sur Facebook. Nous produisons de plus en plus de données. Twitter génère à lui seul 12 téraoctets de données supplémentaires chaque jour, le trafic Internet en une heure pourrait remplir environ sept milliards de DVD (7 fois la hauteur de l'Everest, une fois empilés) ; d'ici 2020, le volume de données généré est estimé à 10,4 zettaoctects, soit 10 400 milliards de gigaoctets de données déversés tous les mois sur Internet et l'univers numérique sera 44 fois plus grand qu'en 2009.

Le Big Data, suscité par des données de plus en plus nombreuses éveille bon nombre d'applications et façonnera tous les secteurs. L'objectif est de retirer une forte valeur ajoutée dans l'analyse des données. Le Big Data peut être utilisé pour détecter les risques sanitaires, la recherche, la cyber sécurité, la recommandation ciblée de produits etc.

Le taux de pénétration très élevé de la téléphonie mobile en Afrique place les données mobiles au cœur de la stratégie Big Data africaine et des pays émergents. Les résumés des communications sont utilisés pour des prévisions et prédictions sur la santé, la mobilité et l'aspect socio-économique.

Le Big Data entraîne avec lui des dépassements de performance au niveau du stockage de données, de la rapidité des analyses et permet aujourd'hui d'agir sur des problématiques analytiques soulevées en temps réel. Les systèmes de gestion de base de données classique sont insuffisants pour analyser des données de plus en plus volumétriques et variées et disparates (4V). De nouveaux outils provenant des géants d'internet vont apporter une solution aux problèmes de stockage et d'analyse ultra rapide. L'informatique en nuage ou cloud computing viendra doper ces nouvelles technologies en rendant l'accès et les configurations plus flexibles.

Les pays émergents devraient surtout utiliser le Big Data afin de répondre aux besoins de développement, de santé et d'économie. Nous pouvons dégager grâce à cette technologie des plans d'actions assez orientés

Toutefois comme toute nouvelle technologie, le Big Data vient avec un lot de préoccupations relatives surtout au respect de la vie privée, à l'espionnage de masse et à une économie de données. Les entreprises souhaitent de plus en plus transformer leurs données en capital mais sont également prêtes à tout pour acquérir les données des autres afin de toujours développer un avantage concurrentiel. Cette course pourrait si l'on n'y prendre garde avoir des conséquences sans précédent.

Raison pour laquelle le rôle du Régulateur n'a jamais été aussi important pour préserver le droit de liberté des citoyens. La Côte d'Ivoire s'est dotée d'une arme juridique, la loi 2013-450 du 19 juin 2013 relative à la protection des données à caractère personnel toujours dans l'optique défendre la vie privée du citoyen, et compte pour cela sur le régulateur pour son application.

Bibliographie

- Kayser-Brill, N. (2015). BIG DATA : Les données changent la donnée. *PARISWORLDWIDE*, 70-82.
- Loi n°2013-450 du 19 juin 2013 relative à la protection des données à caractère personnel
- UNGlobalPulse. (2013). *Mobile phone network data for development.*
- UNGlobalPulse. (2013). *Big Data for development : a primer.* UN.
- Mayer-Schönberger, M. A. (2014). *Big Data, opportunity or threat.* GSR Discussion Paper, ITU-T.
- Adolph, M. (2013). *Big Data, big today, normal tomorrow.* Technology watch, ITU-T, Standardization.
- Deloitte. (s.d.). *Data & Analytics Trends 2015.* Consulté le 02 24, 2015, sur Deloitte-france: http://www.deloitte-france.fr/analytics-trends-2015/home.html
- EMC. (2011, Juin). Consulté le Mars 01, 2015, sur http://www.emc.com/collateral/about/news/idc-emc-digital-universe-2011-infographic.pdf
- IBM. (s.d.). *Bigdata*. Consulté le Février 23, 2015, sur IBM: http://www-01.ibm.com/software/fr/data/bigdata/
- ITU. (2014, Mai 05). *Communiqué de presse.* Consulté le Avril 01, 2015, sur https://www.itu.int/net/pressoffice/press_releases/2014/23-fr.aspx
- Le journal du net. (s.d.). *Big Data.* Consulté le Février 24, 2015, sur Le journal du net: http://www.journaldunet.com/solutions/analytics/big-data/
- Big Data. (2015, Février 08). Consulté le Février 23, 2015, sur Wikipédia, l'encyclopédie libre: http://fr.wikipedia.org/w/index.php?title=Big_data&oldid=11167220
- Commentary: Containing the Ebola Outbreak – the Potential and Challenge of Mobile Network Data. (2014, Septembre 29). Consulté le 04 13, 2015, sur PLOS Current Outbreaks: http://currents.plos.org/outbreaks/article/containing-the-ebola-outbreak-the-potential-and-challenge-of-mobile-network-data/
- Big Data from cheap phones. (2013, Avril 23). Consulté le Avril 13, 2015, sur MIT Technology Review: http://www.technologyreview.com/featuredstory/513721/big-data-from-cheap-phones/
- Ebola. (2014, Septembre 09). Consulté le Avril 13, 2014, sur Worldpop: http://www.worldpop.org.uk/ebola/
- Cloud Computing. (s.d.). Consulté le Avril 13, 2015, sur Wikipédia, l'encyclopédie libre: http://fr.wikipedia.org/wiki/Cloud_computing
- Anonymisation. (s.d.). Consulté le Avril 13, 2015, sur Wikipédia, l'encyclopédie libre: http://fr.wikipedia.org/wiki/

Le service Veille Technologique rattaché à la Direction des affaires Economiques, de la Prospective et de la coopération Internationale (DEPI) de l'ARTCI scrute le paysage des TIC afin de déterminer de nouveaux sujets d'informations. Ces sujets permettent d'analyser l'actualité du secteur, de mieux comprendre les enjeux de la régulation et l'impact des TIC dans la vie de tous les jours

Recherche et Rédaction

ADOPO Antony Virgil
Ingénieur Data Science / IT
DEPI
Téléphone : +225 20 34 58 80 / 80 27
Fax : +225 20 34 43 75
Email : adopo.antony@artci.ci

YAO N'Guessan Kevin
Chef de Service Veille Technologique
DEPI
Téléphone : +225 20 34 58 80 / 80 27
Fax : +225 20 34 43 75
Email : yao.nguessan@artci.ci

Conseil consultatif

ZEBOUA Bagodou Patrick
Chef de Service Etudes et Développement
DEPI
Téléphone : +225 20 34 58 80 / 80 27
Fax : +225 20 34 43 75
Email : zeboua.patrick@artci.ci

COULIBALY Namongo Adama
Chef du Département Prospective Universelle
DEPI
Téléphone : +225 20 34 58 80 / 80 17
Fax : +225 20 34 43 75
Email : coulibaly.namongo@artci.ci

KOUAKOU Guy-Michel
Directeur
DEPI
Téléphone : +225 20 34 58 80
Fax : +225 20 34 43 75
Email : kouakou.guy-michel@artci.ci

ARTCI
AUTORITE DE REGULATION DES TELECOMMUNICATIONS/TIC DE COTE D'IVOIRE

Avril 2015
Abidjan , Côte d'Ivoire