INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES ROUEN

Projet de Physique P6

STPI/P6/2016

BIG DATA

Etudiants :

Anass El YAAGOUBI **Jitao XU**

Chunan JIN **Yizhe WANG**

Elsa LAVAL

Enseignant-responsable du projet :

Abdelaziz BENSHAIR

Date de remise du rapport : ***07/06/2016***

Référence du projet : ***STPI/P6/2016***

Intitulé du projet : ***LE BIG DATA***

Type de projet : ***Culture Scientifique et veille technologique***

Objectifs du projet :

Les objectifs principaux du projet sont la description du phénomène relativement récent que représente le Big Data, et la présentation de méthodes qui permettent sa mise en place. Ce projet a également pour but d'exposer les différentes applications du Big Data dans la société actuelle. Les perspectives de développement ainsi que les limites seront également à traiter. Enfin, le rapport fera la présentation d'un cas d'usage : les Smart Cities.

Mots-clefs du projet :

-Données massives

-Capteurs

-Analyse

-Information

TABLE DES MATIERES

1. INTRODUCTION

Depuis quelques années, les termes « Big Data » font partie intégrante du langage courant. Si cette expression est utilisée de façon abusive et parfois inadaptée, elle illustre un phénomène bien réel. On assiste en effet à la croissance exponentielle du volume de données générées. Selon une étude IBM, 90% des données disponibles aujourd'hui ont été créés au cours des deux dernières années. Ces données représentent une source d'information impressionnante si l'on parvient à les traiter efficacement. Une fois traitées, elles démultiplient les capacités de développement de très nombreux domaines, de la médecine à l'agriculture, en passant par le sport. De plus, les techniques d'analyse prédictive, fondées sur le Big Data, semblent permettre d'anticiper les faits et les actions. La parfaite maitrise de la collecte, de l'analyse et du partage des données apparait alors comme l'un des enjeux de la décennie actuelle.

Les êtres humains sont les principaux émetteurs de ces données. A chaque instant, nous les libérons en utilisant nos ordinateurs, en nous déplaçant avec notre téléphone portable dans la poche, ou encore en faisant nos courses. Cette omniprésence de capteurs, de traceurs de navigation ou même de caméras, pousse certains à s'opposer à ce phénomène. Si l'objectif affiché du Big Data est d'optimiser nos taches et d'améliorer notre vie quotidienne, il permet également à ceux qui détiennent les bases de données de nous ficher et de prévoir nos comportements. Le problème de la capture et de la gestion de nos données personnelles est au cœur des problématiques contemporaines.

Notre choix s'est porté sur le thème du Big Data car nous avions envie de comprendre réellement ce phénomène récent et de découvrir la véritable place qu'il occupe dans la société actuelle. Le but de notre projet a donc été de définir le Big Data de façon précise, et d'en exposer les enjeux. Nos recherches ont servi à couvrir deux axes principaux : le contexte socio-économique ainsi que l'aspect scientifique et technique de l'utilisation des données en masse. Il a fallu veiller à présenter les perspectives d'évolution positives mais également les limites. Enfin, le travail de groupe et la capacité de synthèse sont des compétences qui ont dues être mises en œuvre durant ce projet.

Ce rapport est organisé en trois grandes parties. La première donne une définition du Big Data et en présente l'état de l'art. La seconde consiste en une explication de quelques méthodes permettant la collecte, le stockage et le traitement d'un volume important de données. La dernière partie du rapport est consacrée à l'étude détaillée des applications du Big Data qui rendent possible la création de villes intelligentes.

2. METHODOLOGIE / ORGANISATION DU TRAVAIL

Le travail pour réaliser ce projet consistait principalement à collecter des informations puis à les synthétiser et les organiser entre elles. Nous avons, dans un premier temps, élaboré un plan détaillé. Cela a permis à tout le groupe de visualiser le projet dans son ensemble mais également à la répartition des tâches. En effet, chacun a pu choisir les aspects qu'il préférait approfondir.

La deuxième étape a donc été l'étape de recherche des informations. Nous avons travaillé séparément sur nos propres parties. Nous cherchions de la documentation principalement sur le Web. Les documents sur la notion de Big Data y sont très nombreux. De plus, les encyclopédies en ligne dans le domaine des mathématiques et de l'informatique sont une source d'information fiable. Tous les lundis, nous nous sommes tenus au courant sur l'avancée de nos recherches. Grâce à cela, nous avons pu comprendre les différentes parties du projet et effectuer les modifications nécessaires, au fur et à mesure.

Une fois toutes les informations récoltées, nous nous sommes attaqués à la rédaction. Chacun a rédigé ses paragraphes et nous avons mis en commun. Nous avons ensuite harmonisé nos écrits pour maintenir une cohérence tout au long du rapport. Après avoir terminé le travail sur le rapport écrit, nous avons débuté le travail pour la soutenance.

3. DEFINITION ET ETAT DE L'ART

3.1. Définition générale

C'est en 1997 que l'expression « Big Data » fait sa première apparition, d'abord au sein d'articles scientifiques. Cette notion se répand ensuite rapidement et on la retrouve seulement quelques années plus tard, dans un nombre important de domaines. Les termes Big Data désignent l'explosion du volume de données numériques. Ce volume devient alors impossible à traiter, analyser et stocker à l'aide des outils traditionnels de gestion de base de données. Cette définition étant peu précise, Gartner[1] décrit ce phénomène de façon plus schématique en 2001 : elle le définit comme un regroupement d'outils qui répondent à la problématique dite des 3V.

Le Big Data, c'est traiter des Volumes de données conséquemment supérieurs à ceux traités auparavant, à une Vitesse incomparable, le tout en intégrant une Variété de données largement plus riche.

1 : Volume : le volume des données générées est en constante et rapide augmentation. Chaque jour, nous créons 2,5 quintillions (2 500 000 000 000 000 000) d'octets de données. Les entreprises, dans tous les secteurs d'activité, font appel au Big Data pour gérer ces données qui dépassent les limites des outils auparavant utilisés.

2 : Vitesse : grâce aux évolutions technologiques récentes, les données sont générées et capturées, à très grande vitesse. Le traitement et l'analyse doit donc se faire dans un laps de temps très court, voir infime, pour qu'elles puissent être partagées quelques instants plus tard, ou même en temps réel. Il est évident que cette fréquence engendre un gain en efficacité considérable, ainsi qu'une plus-value économique.

3 : Variété : le Big Data rend possible le traitement de tous les types de données : chiffres, images, sons, vidéos, commentaires. Il n'est plus obligatoire de travailler sur des données très formatées pour garantir la capacité de comparaison.

Il faut savoir que, dans certains cas, on ajoute à ces 3V fondamentaux le V de la Valeur et ou encore celui de Véracité. Ils évoquent la nécessité de posséder de données fiables et pertinentes afin que les analyses réalisées aient du sens.

3.2. Origine des données et acteurs

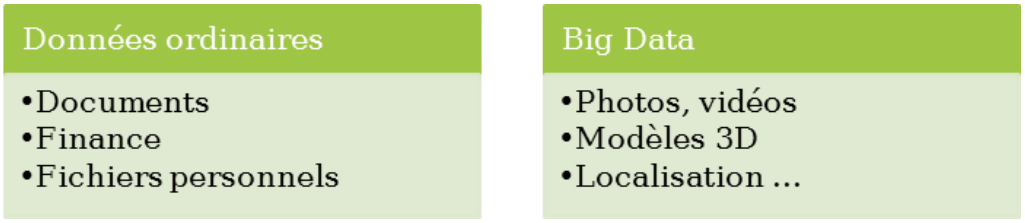
Le Big Data traite des données qui ont des origines très variées. On peut distinguer quatre sources de données principales :

- Le Web
- L'internet et les objets communicants : réseaux de capteurs, téléphonie etc…

[1] Entreprise américaine de conseil et de recherche dans le domaine des techniques avancées.

- Les sciences : climatologie, génomique, astrophysique etc …
- Les données commerciales, publiques ou personnelles.

Les données utilisées grâce au Big Data se démarquent de celles utilisées dans un contexte ordinaire par la diversité de leur format. On retrouve ici la notion suggérée par le « V » de Variété.



C'est sur le Web que ces données hétérogènes sont accessibles le plus facilement. Chaque minute, 240 millions de messages électroniques sont envoyées et 12 heures de vidéos sont postées sur YouTube. C'est pourquoi, ces données disparates ont initialement été récoltées et exploitées par le monde du Web. Les précurseurs du Big Data sont donc les entreprises comme Google, Microsoft, Amazon et Facebook. Ces dernières sont d'ailleurs souvent nommées « Big Four ». Peu à peu, le phénomène s'est répandu, et nombreuses sont les entreprises qui se sont lancées dans l'utilisation massive des données.

La culture et le modèle économique des sociétés du Web ont placé l'Open Source au centre des projets de Big Data. **L'Open Source** correspond au libre accès au code source des programmes informatiques et à l'autorisation de création de travaux dérivés. Le code source résulte lui-même bien souvent d'une collaboration entre des programmeurs. Un des acteurs les plus actifs dans ce domaine est la Fondation Apache. Elle a lancé plus d'une dizaine de projet, dont le plus célèbre est Hadoop.

3.3. Applications :

Les applications du Big Data sont déjà nombreuses et tendent à le devenir de plus en plus. Il est difficile de dresser une liste exhaustive de ces applications, nous allons donc présenter quelques applications majeures mises en œuvre aujourd'hui. Les usages du Big data présentés sont volontairement rattachés à un domaine d'application caractéristique.

- **La santé :** De nos jours, les médecins peuvent diagnostiquer plus précisément les maladies et trouver la meilleure façon de les traiter à l'aide des données massives. Le Big Data permet également d'améliorer et de rentabiliser les recherches et le développement des nouveaux médicaments. On peut aussi construire le dossier médical personnel électronique d'un patient et prédire son état de santé en possédant d'innombrables informations sur ses habitudes de vie.

- **Le commerce :** Les grands magasins, comme par exemple Walmart et Macy's, traitent plus d'un million de transactions par heure. Ces données sont utilisées afin d'obtenir des informations sur leurs produits et d'en adapter les stocks et les prix. Les entreprises de e-commerce, qui suivent le modèle d'Amazon, utilisent aussi le Big Data, pour mieux connaitre le client et cibler ses besoins.

- **Le sport :** On applique le Big Data dans le domaine du sport pour une meilleure connaissance des athlètes grâce aux capteurs, dans le but de leur fournir des entrainements encore plus adaptés. Ces données sont aussi utilisées par les équipementiers,

qui cherchent à perfectionner les tenues et le matériel sportif. Une autre application du traitement de bases de données gigantesques est liée au sport : le marché de paris sportifs. Grâce aux techniques d'analyse prédictive *(présentées en 3.4)*, il est désormais possible, en étudiant énormément de paramètres, de pronostiquer avec beaucoup de précision le gagnant d'un match. Toutefois, cette méthode n'est pas encore entièrement fiable, les prédictions restent parfois erronées. Les calculs n'ont pas effacé l'idée des « aléas du sport ».

Figure 1: Estimation du nombre de données générées[2]

- **Les infrastructures intelligentes** : Les bâtiments ont de multiples systèmes qui produisent des données : de la gestion de l'infrastructure, qui capte les températures et le taux d'humidité par exemple, au contrôle de l'accès, qui recueille les statistiques d'occupation. Le Big Data nous fournit une plate-forme ouverte, qui permet l'intégration des données provenant de ces différents systèmes. Une fois combinées et normalisées, les opportunités du Big Data peuvent alors émerger. En suivant les chauffages, les climatiseurs, les ventilateurs et les lumières, on va accumuler une très grande quantité des données. Les analyses et les algorithmes peuvent être exécutés sur ces données pour parvenir à faire des économies d'exploitation et d'énergie. La détection des défauts et l'optimisation de la construction sont les deux méthodes principales pour faire des économies, tout en fournissant un retour rapide du coût de l'installation de la solution.

3.4. L'analyse prédictive

Nous allons maintenant exposer de façon plus détaillée une des plus applications essentielles du Big Data: l'analyse prédictive.

L'analyse prédictive est un ensemble de techniques issues des probabilités et des statistiques. Elle permet d'extraire des informations et des modèles en se basant sur des données du passé, dans le but de faire des hypothèses prédictives sur des événements du futur. Elle est mise en œuvre grâce Big Data. En effet, depuis l'utilisation des données massives, la portée de l'analyse prédictive se développe de façon exponentielle.

Il existe différentes méthodes d'analyses prédictives:

- **Les Modèles prédictifs** permettent, à partir de données antérieures, de prédire un événement futur plus ou moins probable afin d'améliorer l'efficacité dans une branche donnée (par exemple, la prévision boursière).
- **Les Modèles descriptifs** étudient des relations entre les données pour classer des individus ou des produits dans des groupes.
- **Les Modèles de décision** sont utilisés pour identifier un ensemble de règles permettant de définir une logique de décision.

[2] Donnée fournie par les responsables de l'équipe nationale allemande de football

Ces modèles permettent de prendre des décisions plus précises, de façon rapide et peu coûteuse. Ils améliorent nos choix en augmentant leur efficacité et leur rentabilité.

L'analyse prédictive couvre un vaste champ d'applications. Par exemple, elle est utilisée en finance où elle est devenue indispensable notamment dans les marchés boursiers, mais aussi en assurance afin d'évaluer les risques, en télécommunications, en marketing, en médecine, en météorologie et bien d'autres domaines.

Après cet état de l'art du Big Data, il semble intéressant de rentrer plus en détail sur l'aspect scientifique et technique de sa mise en place. Nous verrons comment les données sont capturées, traitées puis stockées et émises.

4. RECUEIL, STOCKAGE ET TRAITEMENT DES DONNEES

Les capteurs font partie des éléments qui participent à la naissance du Big Data. Outre les données Web, la majorité des autres données en masse sont récoltées à l'aide de capteurs. Sans cette source considérable de données, les enjeux du Big Data seraient moindres.

On les retrouve partout, dans nos maisons, sur la route pour aller au travail, dans nos bureaux, dans les hôpitaux etc… Ils ne nous quittent plus, surtout avec l'arrivée des objets connectés. Ils sont devenus omniprésents, ils enregistrent les données en temps continu à travers les objets communicants. Ces objets remplis de capteurs sont innombrables et leur nombre ne cesse d'augmenter. Citons par exemple la brosse à dents qui nous dit pendant combien de temps on s'est brossé des dents et combien de fois par jour, ou encore le pacemaker qui enregistre l'activité cardiaque et qui est constamment connecté à internet pour envoyer les données au médecin du patient.

4.1. Définitions

Le capteur est un dispositif permettant de transformer une grandeur physique en une grandeur manipulable facilement par exemple une tension électrique qu'on pourra éventuellement transmettre et enregistrer par la suite.

Voici une liste non exhaustive de grandeurs physique pouvant être enregistrées par un capteur dans un ordre alphabétique:

1. Angle
2. Contrainte
3. Courant
4. Champ magnétique
5. Débit
6. Force
7. Inertie
8. Lumière
9. Niveau
10. Position
11. Pression
12. Son
13. Température

Le nombre de catégories de capteurs est important, c'est pourquoi nous avons choisi d'en étudier précisément une seule. Parmi ces grandes catégories de capteurs, nous allons nous intéresser à celles des capteurs du trafic routier, car ces derniers reposent sur des principes physiques très divers. De plus, ces capteurs sont largement utilisés dans les projets de Smart City que nous avons décidé d'étudier. Ils sont classés, de façon générale, en quatre grandes familles:

4.1.1. Capteurs intrusifs

(a) Les capteurs à boucles inductives
(b) Les capteurs « magnétomètres » en chaussée
(c) Les capteurs à effet « Piézo-électrique »
(d) Les capteurs à « Jauges de contraintes »
(e) Les capteurs à Tubes pneumatiques
(f) Les capteurs Résistifs
(g) Les capteurs à « Fibres optiques »

4.1.2. Capteurs non-intrusifs

(a) Les capteurs hyperfréquences utilisant l'effet Doppler

(b) Les capteurs hyperfréquences utilisant deux antennes

(c) Les capteurs Laser

(d) Les capteurs à Infrarouge actifs et passifs

(e) Les capteurs Vidéo visibles et infrarouges

(f) Les capteurs vidéo spécialisés dans la Lecture Automatique de Plaques d'Immatriculation (LAPI) visible et infrarouge

(g) Les capteurs spécialisés dans la Détection Automatique des Incidents (DAI)

(h) Les capteurs acoustiques passifs (microphone)

(i) Les capteurs à « Couplage de technologies »

(j) Les capteurs acoustiques actifs (à Ultrasons)

4.1.3. Capteurs embarqués

4.1.4. Systèmes coopératifs et recueil de données

(a) Systèmes autonomes

(b) Systèmes coopératifs

Les Capteurs intrusifs sont posés dans la chaussée d'où le nom intrusif car ils déforment la chaussée. Les Capteurs non-intrusifs sont des capteurs qui ne nécessitent pas un travail de la chaussée.

Dans la suite nous allons nous intéresser aux capteurs piézo-électriques et aux capteurs à effet Doppler pour illustrer les types intrusifs et non-intrusifs. Cependant on ne va pas développer les autres types de capteurs. S'ils possèdent des caractéristiques propres, une liste exhaustive de tous les capteurs existants serait longue et complexe et n'apporterait rien à ce rapport.

4.2. Fonctionnement de deux types de capteurs :

4.2.1. Effet piézo-électrique

L'effet piézo-électrique est un phénomène électromécanique découvert par les frères Curie en 1880. Ce phénomène est une propriété qu'ont certains cristaux (comme le Quartz) à se charger et décharger lorsqu'une force orientée convenablement est appliquée au matériau. Le schéma de la figure 2 ci-après le montre.

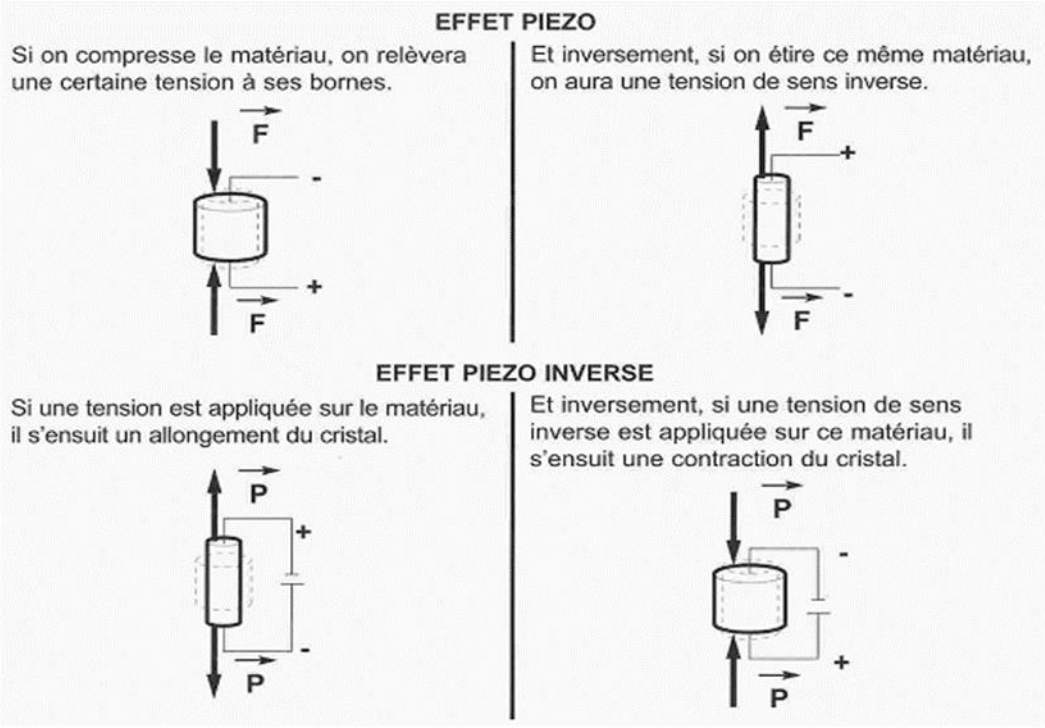


Figure 2 : Explication détaillée de l'effet piézo-électrique

Si une force orientée convenablement est appliquée au matériau, il apparait alors une tension aux bords du matériau proportionnelle à l'effort appliqué. C'est ce principe (entre autres) qui est utilisé pour savoir si une place de parking est libre ou non, ou pour mesurer le trafic sur une portion de route. On a constaté que l'idée jaillissait souvent d'utiliser ce principe pour générer de l'électricité : on peut citer l'exemple d'un stade de foot à Rio de Janeiro éclairé grâce à l'énergie des joueurs.

A l'inverse, si une tension électrique est soumise à un matériau possédant des propriétés piézoélectriques, on observe alors une déformation de celui-ci. C'est ce que l'on appelle l'effet piézo inverse. Ce principe est très utilisé dans les nanotechnologies notamment dans les minis moteurs piézoélectriques.

4.2.2. Description du capteur piézo-électrique :

Le capteur piézo-électrique le plus utilisé dans le trafic routier est à base de céramique. Ce capteur est constitué d'un câble coaxial comportant une gaine et une âme conductrice en cuivre. Ce câble est inséré de manière transverse dans la chaussée, il est enrobé dans un barreau de résine pour qu'il conserve sa rigidité. La force d'impact des pneus d'un véhicule provoque une tension électrique dans la céramique qui est mesurée entre l'âme et la gaine.

Ce type de capteur donne accès à des mesures comme le poids des véhicules, le nombre de véhicules traversant la chaussée, la vitesse, la distance inter-essieux et la position des véhicules sur la chaussée. Il donne également une idée sur le type de véhicule : poids lourd ou non.

4.2.3. L'effet Doppler

L'effet Doppler, et plus précisément l'effet Doppler-Fizeau, est un phénomène présenté en 1842 par le physicien autrichien Christian Doppler, et confirmé par la suite par Hippolyte Fizeau (physicien français du XIXème) pour les ondes électromagnétiques. Buys Ballot le confirme enfin pour les ondes sonores, grâce à une célèbre expérience consistant à faire jouer une note calibrée à des musiciens sur un train. Quand le train s'approchait les spectateurs de la gare entendaient une note plus aigue et quand le train s'éloignait ils entendaient une note plus grave.

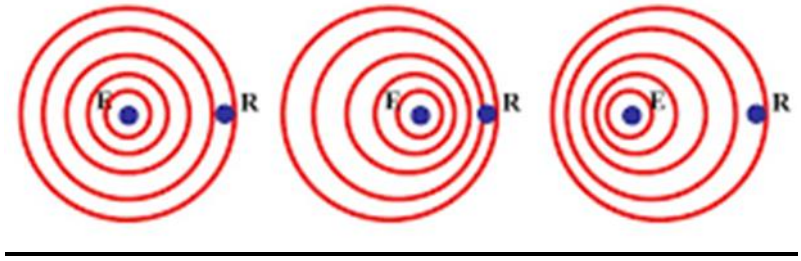


Figure 3: Illustration de l'effet Doppler

i) Lorsque la source est immobile la fréquence perçue en R est la même qu'en E.

ii) Lorsque la source s'approche de R la fréquence perçue en R est plus grande qu'en E.

iii) Lorsque la source s'éloigne de R la fréquence perçue en R est plus faible qu'en E.

Si on ne tient pas compte des phénomènes relativistes, on peut résumer l'effet Doppler à l'aide de la formule suivante:

$$f = \frac{c - v}{c - ve} fe$$

où *ve* est la vitesse d'émetteur, *v* est la vitesse du capteur, *c* est la célérité de l'onde, *fe* est la fréquence d'émission et *f* est la fréquence perçue. Dans le cas du Radar immobile, on a *v=0* et par conséquent $f = \frac{c}{c-ve} fe$ qui après transformation devient $= \frac{f-fe}{f} c$. On retrouve bien les trois remarques précédentes...

On a constaté que les capteurs utilisant l'effet Doppler sont utilisés dans différents domaines très variés de la médecine à l'astronomie en passant les capteurs routiers...

4.3. Le stockage des données récoltées

Il est relativement simple de récolter des données, mais pour pouvoir les utiliser il faut d'abord les enregistrer c'est à dire les stocker. Ainsi le stockage des données est un challenge de taille. En effet, la croissance des données est exponentielle et ce sont plus de deux trillions d'octet qu'il faut stocker tous les jours. Pour cela des Data Centers ont été mis en place.

Ces centres de données peuvent atteindre des tailles immenses (60 000m^2), gérer des centaines de milliards d'Octets et consommer autant que des villes de centaines de milliers d'habitants.

4.3.1. Les Data Centers

Un Datacenter ou centre de données est un bâtiment regroupant des équipements électroniques tels que des serveurs qui sont des ordinateurs très puissants, servants pour les télécommunications, le traitement et le stockage des données. Son usage principal reste tout de même le stockage de l'information. Les petites comme les grandes, toutes les entreprises ou presque y ont recourt, parfois de façon indirecte par exemple pour l'hébergement de leur site internet. Les entreprises de grande taille quant à elles font souvent appel à plusieurs centres de données afin de sauvegarder les données de leurs clients. On peut citer les banques, les grands sites d'achat en ligne, certains centres de recherche scientifique et tous les réseaux sociaux Facebook, Twitter…

A. Description :

Un Datacenter n'est pas seulement un tas de serveurs dans un endroit quelconque. Au contraire c'est un ou plusieurs complexes fonctionnants ensemble, sur des dizaines voire des centaines de milliers de mètres carrés, contenants des milliers de serveurs.

Ils sont organisés en baies et chaque baie mesure de 60 à 80 cm de large et plusieurs mètres de hauteur. Elle peut contenir jusqu'à 80 ordinateurs (qui sont composés de processeurs, de disque durs..), c'est ce qu'on appelle serveurs.

Les centres de données sont des endroits très protégés car ils peuvent contenir des données très sensibles, comme des e-mails ou des données de transactions financières. Un autre challenge des centres de données est le refroidissement des serveurs car ils produisent beaucoup de chaleur. Il y a aussi un souci d'optimisation de leur consommation électrique, car ces derniers sont très gourmands en énergie. Afin d'alimenter son million de serveurs le géant Google aurait consommé près de 300 MW d'électricité, ce qui représente l'équivalent de la consommation électrique d'une ville comme Marseille.

B. Support d'information électronique:

Les technologies de stockage de l'information évoluent de manière très rapide, ainsi l'humain a commencé par l'utilisation de la gravure sur les pierres puis des parchemins. Il y a encore deux siècles, on utilisait un support papier pour stocker l'information, mais au XX^{e} siècle l'arrivée du numérique révolutionna ce stockage : en effet une clé USB de 32 GO peut contenir environ 60 longs métrages en format MPEG.

On peut distinguer deux types de mémoire: l'une est dite de masse (grande quantité et long terme) et l'autre dite rapide (qui est une mémoire de travail).

I. Stockage de masse :

1. Première génération
   i. La carte perforée
   ii. Le ruban perforé.
2. Deuxième génération
   i. La bande magnétique
   ii. La cassette
   iii. Le disque dur
   iv. La disquette
3. Troisième génération
   i. Le disque compact
   ii. Le DVD (Digital Versatile Disc)
   iii. Le Blu-ray.
4. Quatrième génération
   i. La clé USB (Universal Serial Bus)
   ii. La carte SD (Secure Digital)
   iii. La Carte microSD

II. Stockage à accès rapide: mémoire de travail:

a. Mémoire vive RAM: Random Access Memory= mémoire volatile.

b. Mémoire morte ROM: Read Only Memory = une mémoire non volatile.

Un disque dur, parfois abrégé HDD (Hard Disk Drive,) est un dispositif de stockage de masse, qui est essentiel aux ordinateurs (entre autres). La première version de disque dur fut construite par IBM en 1956, elle s'appelait IBM 350 et fut utilisée dans le RAMAC 305(Ordinateur).

Les disques durs de nos jours sont composés de quatre éléments principaux : de plateaux où est stockée l'information, d'un axe de rotation, de têtes de lecture/écriture pour chaque surface de plateau et d'un système électronique intégré permettant son fonctionnement[3].

- Les plateaux sont en liaison (roulements à billes) avec l'axe de rotation. Dans le passé, ils étaient composés d'aluminium ou zinc. De nos jours ils sont de plus en plus composés de verre et traités par diverses couches. Une de ces couches est ferromagnétique c'est elle qui stocke l'information. Elle est revêtue d'une dernière couche de protection. Les plateaux sont subdivisés en secteurs. Chaque secteur est identifié par un numéro de piste, un numéro de tête et un numéro de secteur.
- L'axe de rotation est entrainé par un moteur à une vitesse de rotation constante (par exemple : 7200 tr/min).
- Les têtes de lecture/écriture sont fixées au bout d'un bras mobile leur permettant de pivoter en arc de cercle survolant ainsi la surface du plateau sans le toucher à une distance d'environ 10 nm. En survolant le plateau, la tête mesure une résistance dépendant de l'orientation du champ magnétique sur une petite région : c'est la lecture du bit associé à cette région. L'écriture quant à elle est assurée par une tête inductive c'est la technologie GMR (Giant Magneto Resistance). Lorsque la tête touche le plateau cela est appelé atterrissage et cause la destruction de l'information située à cet endroit.
- Le bras mobile est actionné par une bobine parcourue par un courant électrique, (force de Laplace) permettant de contrôler le bras et par conséquent la position des têtes.

L'évolution des techniques liées au stockage de l'information peut se résumer en quatre objectifs:

- Plus de capacité
- Plus de rapidité
- Plus de fiabilité
- Moins de coûts

Depuis le milieu du XXème la performance et la capacité des supports de l'information n'ont cessé d'augmenter de façon exponentielle. La loi de More affirme que le nombre de transistors double tous les deux ans.

[3] Voir annexe 1

4.4. Le traitement des bases de données

S'il est, d'un point de vue technique, compliqué de recueillir et de stocker un nombre important de données en peu de temps, la tâche la plus complexe est de procéder à leur traitement afin d'en tirer des informations pertinentes. Ceci est à la fois difficile à étudier théoriquement et à mettre en œuvre. Il faut que les technologies utilisées permettent de valoriser les données brutes et hétérogènes. Un des plus grands enjeux du Big Data réside ainsi dans la maitrise de nouvelles techniques d'analyses de données. Il est ainsi intéressant de comprendre les outils mathématiques et informatiques utilisés.

4.4.1. Statistiques et analyse

Cette première étape analyse les données de façon traditionnelle. C'est-à-dire que la conclusion de cette étape ne fournit pas une prévision précise, mais seulement un résultat d'analyse des caractères des données ou des relations entre les événements. De plus, elle sert à classifier et extraire simplement des données.

Il y a beaucoup de techniques très connues et largement utilisées pour effectuer cette étape : le Test d'hypothèse, le Test de Student, l'analyse de la variance, le Test du X2 pour l'analyse des caractères de données, la régression, l'analyse factorielle, l'analyse discriminante, ou encore l'analyse factorielle des correspondances pour chercher les relations entre les événements. On peut également ajouter le partitionnement de données pour la classification et l'analyse en composantes principales pour extraire des données de façon simple. Plusieurs compagnies rendent le service de faire ces calculs statistiques et ces études sur des données. Par exemple, on peut citer GreenPlum de EMC, Exadata de Oracle, infobright de MySQL et Hadoop.

4.4.2. Exploration des données

Cette seconde étape permet des analyses plus précises. C'est-à-dire que l'on peut obtenir une conclusion sur une estimation ou une prédiction. Les outils les plus importants pour traiter des données à ce stade sont les algorithmes plutôt complexes, qui sont utilisés dans les programmes. On utilise différents algorithmes pour satisfaire les demandes. Les demandes plus courantes sont les suivantes :

-Classification

-Groupement d'affinité

-Règles d'association

-Regroupement

-Estimation

-Prédiction

Nous allons maintenant voir quelques exemples d'algorithmes liés aux trois premiers besoins :

A. Classification :

Un classificateur est un outil qui extrait des données en prenant un groupe de données représentant les variables que l'on veut classifier et en tentant de prévoir la classe à laquelle les nouvelles données vont appartenir.

C4.5 est un algorithme qui construit un classificateur sous la forme d'un arbre de décision. Il fait une classification en posant une question après l'autre pour chacune des données, et dès qu'il trouve la réponse, il place la donnée dans une des deux classes qu'il crée. C4.5 possède un ensemble de données qui représentent les éléments déjà classés. Dans un même genre de fonctionnement, on peut mentionner les **Support Vectors Machines** (SVM). Ils sont similaires à C4.5 mais n'utilisent pas les arbres de décision. Ils réalisent la classification à l'aide de vecteurs. Il cherche l'hyperplan les données en deux classes dans un espace à n dimensions. C'est exactement ce qu'une ligne fait dans un espace à 2 dimensions, ou un plan dans un espace à 3 dimensions.

Il existe une famille d'algorithmes qui permet également de classifier en utilisant un autre procédé : **NaiveBayes.** C'est une famille d'algorithmes de classification qui fonctionne tous suivant la même hypothèse : la probabilité de chaque événement est indépendante de celles autres événements de la même classe (deux événements sont indépendants lorsque la valeur d'un événement n'a pas d'influence sur la valeur de l'autre événement).

B. Groupement d'affinité ou de règles d'association

L'algorithme **Apriori** est un algorithme d'apprentissage de règles d'association appliqué à une base de données qui contient un grand nombre de transactions. Il sert à trouver les ensembles qui apparaissent fréquemment dans une base de données et à en déduire une classe. L'apprentissage des règles est une technique d'exploration de données pour connaitre les corrélations et les relations entre les variables dans une base de données. Cet algorithme trouve en premier des sous-ensembles qui ont un cardinal de un et qui apparaissent le plus de fois, et puis cette opération est répétée pour chaque cardinal d'ensemble jusqu'à atteindre la taille définie précédemment.

C. Regroupement

On utilise un programme nommé **K-means**. Il crée k groupes à partir d'un ensemble d'objets de sorte que les membres d'un groupe soient le plus semblables possible. C'est une technique populaire d'analyse de cluster pour explorer un ensemble de données.

L'analyse de cluster est une famille d'algorithmes conçus pour former des groupes tels que les membres du groupe soient plus semblables que les membres du non-groupe. Les mots clusters et groupes ont le même sens dans le domaine de l'analyse de cluster. Il choisit k points en représentant la position moyenne des k partitions initiales. Ensuite, il met chaque point dans la partition dont la position moyenne est la plus proche. Après, à partir des données dans chaque partition, il en calcule une nouvelle position moyenne. Et il ensuite répète les deux étapes. Enfin, il se stoppera si les positions moyennes ne changent pas, c'est-à-dire lorsque la répartition de données ne change plus.

4.4.3. Outils

Les outils utilisés pour traiter ces données sont divers et complexes. Dans le cadre de cette étude, nous considérons qu'il n'y a pas de véritable intérêt d'en faire une présentation détaillée. Nous pouvons tout de même retenir le nom de l'un des outils le plus utilisé : **Mahout** de **Hadoop.**

5. CAS D'USAGE : LES SMART CITIES

Une fois la définition précise donnée et le fonctionnement du Big Data énoncé, nous allons maintenant faire l'étude plus détaillée d'un cas d'usage lié à notre vie quotidienne. Nous avons choisi de présenter les villes intelligentes (ou Smart Cities) comme application du Big Data. Elles mettent bien en avant la polyvalence de ce phénomène. En effet, la maitrise des données massives ouvre beaucoup de possibilités en ce qui concerne l'amélioration de la qualité de vie dans une ville. Récolter et analyser de nombreuses de données peut permettre d'améliorer :

-la qualité de l'air

-le service de transport

-la gestion des déchets

-la consommation d'énergie

-la prévention des crimes …

Figure 4: Caractéristiques d'une ville intelligente

Si le nombre de villes qui utilisent le Big Data est encore relativement faible, notamment en France, ce nombre est amené à augmenter considérablement ces prochaines années. Nous allons donc étudier divers usages des données massives par les villes précurseurs, qui peut être, se généraliseront à la majorité des grandes villes d'ici 2020.

5.1. Exemples d'application dans les villes intelligentes

5.1.1. La gestion du trafic routier

A Birmingham, de nombreuses informations sont transmises à un centre de contrôle pour régler le trafic en temps réel. On utilise les capteurs et les caméras installés sur la route et les feux pour obtenir des informations sur le trafic. A Barcelone, les feux intelligents de circulations sont également utilisés. Ils sont connectés au réseau de trafic. Chaque capteur donne un paramètre différent de flux de circulation. Le système prend des décisions selon les valeurs et donne les instructions adaptées aux feux et aux signaux. Plus le système reçoit de

données, plus il va donner des instructions précises. Ainsi dans cette grande ville d'Espagne, en cas d'urgence, tous les feux qui sont proches de la voiture d'urgence deviendront verts pour que les services d'urgences puissent arriver à temps.

5.1.2. Smart Grid : le réseau d'électricité intelligent

Les équipements utilisés pour mettre en place les Smart Grids sont des capteurs qui reçoivent des données au cours de la production, la transmission, la distribution et la consommation de l'électricité. A l'aide d'une l'analyse sur ces données massives, il possible de localiser rapidement des pannes par exemple. Le système d'analyse permet aussi de mettre en place un prix dynamique de l'électricité. C'est une façon pratique de lisser les pics de consommation en appliquant des charges élevées pendant les heures de pointes et des charges plus basses le reste du temps. De plus, ce système fournit des informations d'usage d'énergie presque en temps réel aux consommateurs, qui peuvent alors gérer leur utilisation sur la base de leurs besoins tout en connaissant leurs charges. Le Smart Grid améliore réellement l'efficacité, la fiabilité, l'économie et la durabilité de la production et de la distribution de l'énergie électrique au sein d'une ville. Il faut savoir que la France est le premier investisseur européen concernant les Smart Grids, avec quelques 118 projets, représentants un investissement d'environ 500 millions d'euros.

5.1.3. Gestion des données statistiques

En 2009, le gouvernement américain a lancé le site internet data.gov, regroupant au départ 47 bases de données. C'est un entrepôt qui en contient aujourd'hui 420,894 couvrant les données des USA sur transport, l'économie, les soins de santé, l'éducation et bien d'autres domaines. Cette initiative représente une étape vers l'augmentation de la transparence, et de l'accessibilité des données au public. Elle est en lien direct avec la promotion de l'Open Data, fortement liée au Big Data. En 2011, Syracuse, une ville de l'État de New York, collaborant avec IBM, a lancé un projet Smarter City qui consistait à utiliser les grandes données pour résoudre ces problèmes de propriétés vacantes. Grâce à l'analyse des données, la ville a été en mesure de cibler 2000 logements vides à récupérer.

5.1.4. La prévision policière : PredPol

Une autre façon d'améliorer la qualité de vie au sein d'une zone urbaine est de maintenir l'ordre. La garantie de sécurité est essentielle pour la vie quotidienne des habitants. Malgré tout, garantir la sécurité est une tâche complexe. Certaines villes ont utilisé le Big Data pour diminuer les délits, petits ou grands, notamment grâce à l'analyse prédictive. La prévision policière fait référence à l'usage des mathématiques et aux techniques de prévision afin de prédire des activités criminelles potentielles. PredPol est un algorithme de prévision policière développé aux Etats Unis afin de prévoir les délits avant qu'ils ne se produisent. Il s'appuie sur une base de données gigantesque, contenant toutes les informations sur les crimes ou infractions antérieures. C'est devenu une pratique courante aux Etats Unis : PredPol a déjà été adopté en Arizona, Californie, Caroline du sud, Illinois, Tennessee et Washington mais également en Angleterre et en Chine. Cet algorithme a fait ses preuves : il a par exemple permis de réduire les cambriolages de 19% à Santa Cruz en Californie.

.

5.2. Les problèmes d'éthique

Ces capteurs qui nous suivent partout et ces caméras qui nous filment à chaque instant rendent possible de belles avancées technologiques. Cependant ils pénètrent dans le même temps dans notre vie privée, parfois à notre insu. Quotidiennement, des milliers de données sur chacun d'entre nous sont enregistrées : données GPS de localisation, calories perdues, heures de sommeil, courses de taxi, ou encore achats et transactions financières. Avec les méthodes que l'on connait aujourd'hui, les détenteurs de ces données personnelles peuvent nous ficher et anticiper nos comportements. C'est d'ailleurs la marque de fabrique des 4 géants du Web, qui utilisent sans limite l'analyse prédictive afin par exemple de cibler notre publicité.

Face à cela, certaines personnes expriment leur mécontentement. Ce fut le cas par exemple lors du scandale sur le programme de surveillance électronique PRISM, relevant de la NSA. En juin 2013, Edward Snowden affirme que la NSA possède un accès direct aux données hébergées par les géants du Web. Cela s'apparente alors à de l'espionnage. Cette révélation avait engendré de nombreuses manifestations aux Etats-Unis.

Par ailleurs, la gestion des données personnelles est également un combat au quotidien, et de multiples lois tentent de régir leur collecte et leur utilisation. En France, la Commission Nationale de l'Information et des Libertés se charge de veiller au respect de la vie privée. Mais la CNIL, créée avant l'apparition du Big Data, n'est plus pleinement adaptée et doit mettre à jour ses méthodes.

Avec le Big Data, la nécessité de bien gérer les données personnelles s'est accélérée et se révèle comme l'un des enjeux majeur de ce siècle. Il faudra donc veiller à créer des technologies performantes mais également rigoureuses par rapport à notre vie privée.

6. CONCLUSIONS ET APPORTS PERSONNELS

6.1. Conclusion de l'étude

Cette étude sur le Big Data nous a confirmé sa place dans la société actuelle, et elle nous a éclairé sur les enjeux qui lui sont liés. Si nous avons observé que tirer des informations sur des volumes de données gigantesques est une tache compliquée, les hommes cherchent tout de même à maitriser cette compétence. En effet, les possibilités qu'offre le Big Data sont si nombreuses et attrayantes qu'il est difficile d'envisager une société future qui s'en passerait. Pour autant, l'évolution du Big Data est freinée par des problèmes d'ordre éthique. Comme pour toute avancée technologique, il faudra veiller à former les utilisateurs du Big Data ainsi que toute la société à ce nouvel univers des données en masses.

6.2. Apports personnels

Outre le travail sur le Big Data, cet E.C projet nous a également permis de découvrir le monde de la recherche en informatique et en mathématiques. Notre encadrant nous a présenté le travail mené par son équipe sur le véhicule intelligent, et nous avons visité le laboratoire du LITIS. En découvrant le sujet de quelques thèses de ce laboratoire, nous avons pu nous construire une meilleure idée de ce que représente la recherche dans ce domaine. Nous regrettons tout de même l'annulation de la visite à l'Institut National de Recherches et Informatique et en Automatique (INRIA). Nous avons joué de malchance.

Malgré tout, Fawzi Nashashibi, chercheur du département RITS[4] de l'INRIA, a eu la gentillesse de se déplacer pour pallier cette annulation. Il nous a présenté de cet institut de recherche international qui compte pas moins de 3500 chercheurs, dont quelques 1000 doctorants : une belle illustration de l'importance des métiers de la recherche. Il a exposé de multiples travaux réalisés à l'INRIA, dans le domaine du véhicule autonome. Nous avons pu apprécier la diversité de ces travaux, liés aux objectifs suivants :

- Amélioration des techniques d'assistance à la conduite
- Création de véhicules autonomes (personnels ou de transport de masse)
- Meilleure gestion du trafic routier

Cette présentation a été également profitable pour notre sujet sur le Big Data. La majorité des technologies en cours de développement qui ont été présentées, nécessitent l'utilisation de données en masse. Cela nous a convaincu de l'omniprésence du Big Data dans les technologies du futur.

Finalement, cet EC et les différentes présentations ont participé à la construction de notre projet professionnel.

[4] Robotics and Intelligent Transportation Systems

7. BIBLIOGRAPHIE ET TABLES

7.1. Articles de périodiques :

« Montpellier teste le Big Data au service de la ville ». Le Monde, 12 avril 2016

« Pour la sociologue Evelyn Ruppert, il faut un regard autre que technique sur le Big Data ». L'Usine Nouvelle, 25 avril 2016

7.2. Publications scientifiques :

Thèse de Lauric Garbuio sur la Lubrification électroactive.

7.3. Sites Internet :

http://future.arte.tv/fr/big-data-la-serie, visité le 10 février 2016.

http://jisajournal.springeropen.com/articles/10.1186/s13174-015-0041-5, visité le 2 mars 2016.

http://www.equipementsdelaroute.equipement.gouv.fr, visité le 12 avril 2016.

http://philippe.boursin.perso.sfr.fr/pdgdies3.htm, visité fin avril 2016.

http://rayli.net/blog/data/top-10-data-mining-algorithms-in-plain-english, visité le 12 mai 2016.

http://link.springer.com/article/10.1186/s13174-015-0041-5, visité le 22 mars 2016.

7.4. Encyclopédies en ligne :

https://fr.wikipedia.org/wiki/Capteur, visité le 12 avril 2016.

https://fr.wikipedia.org/wiki/Effet_Doppler, visité début mai 2016.

https://fr.wikipedia.org/wiki/Stockage_d%27information, visité le 8 mai 2016.

https://en.wikipedia.org/wiki/Data_center, visité le 8 mai 2016.

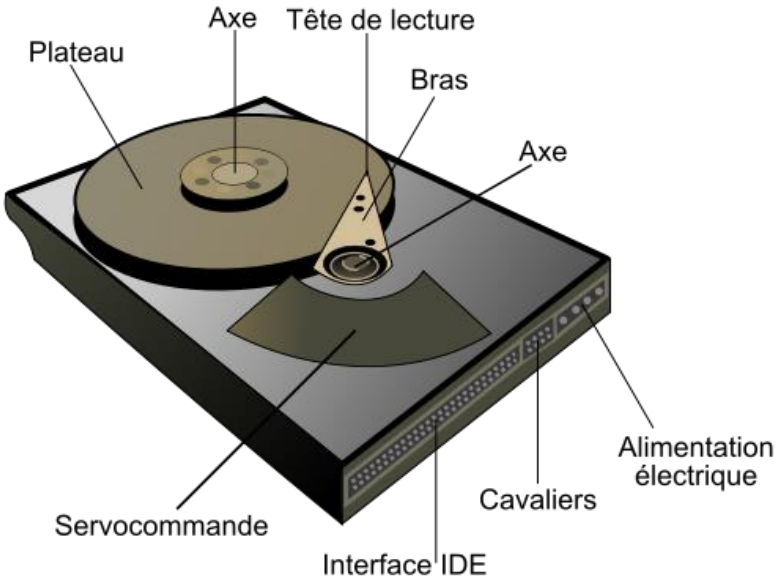
https://en.wikipedia.org/wiki/C4.5_algorithm, visité le 26 avril 2016.

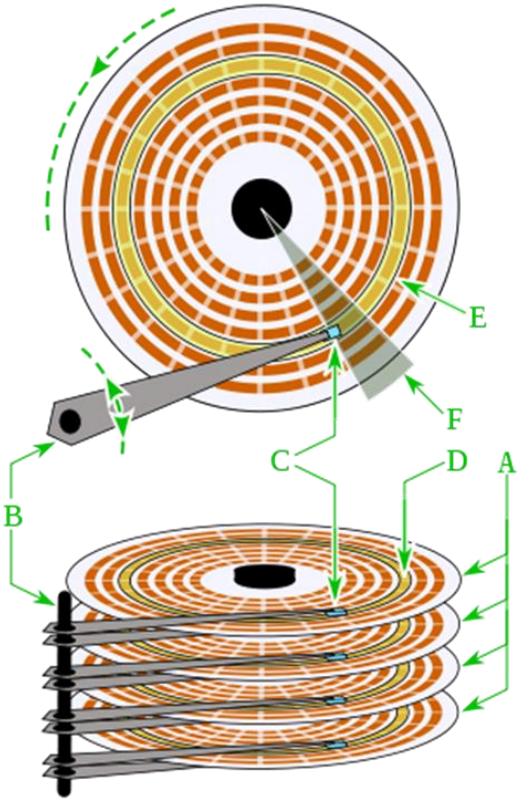
https://fr.wikipedia.org/wiki/Algorithme_APriori, visité le 26 avril 2016.

8. TABLE DES ILLUSTRATIONS

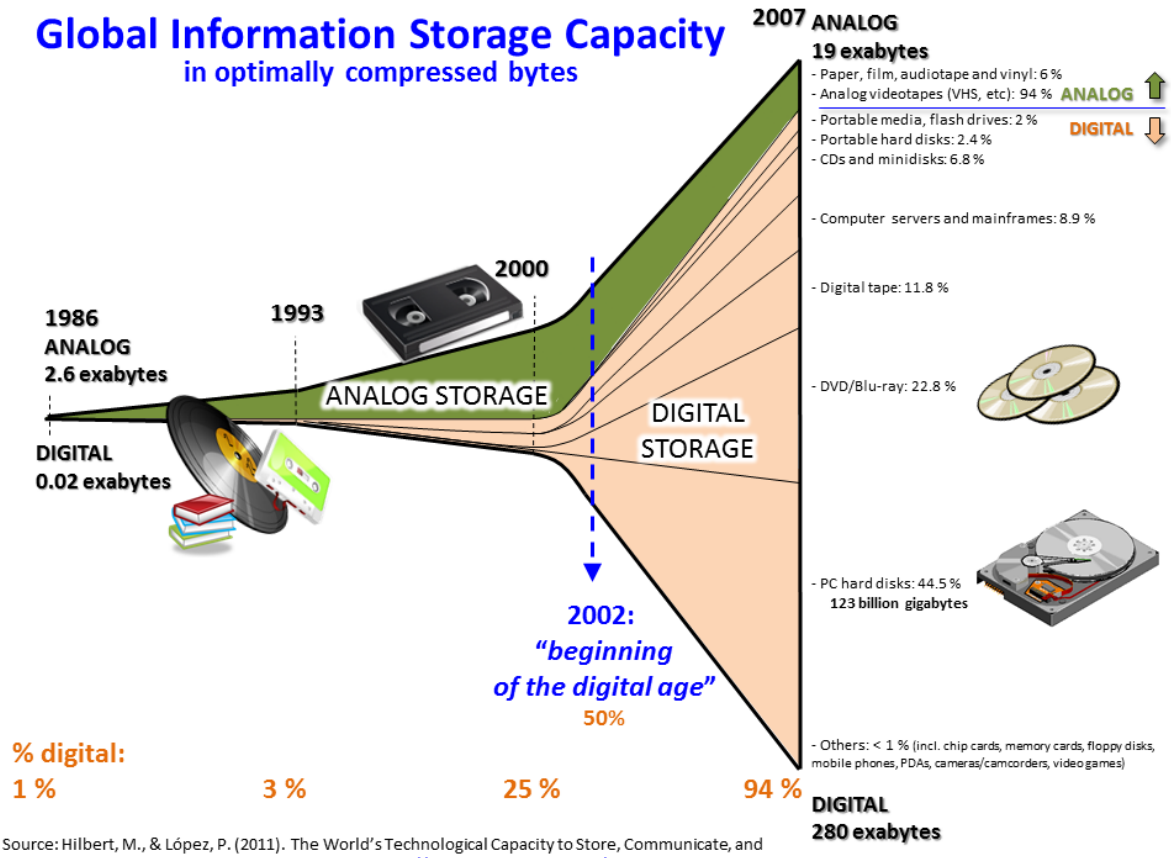
9. ANNEXES

9.1. Annexe 1 : Schéma et fonctionnement d'un disque dur





9.2. Annexe 2 : l'évolution du stockage des données



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html