Université Sidi Mohamed Ben Abdellah

Faculté des sciences Dhar El Mahrez Fès

Master Big Data Analytics & Smart Systems

BIG DATA & ALGORITHMES

2018 - 2019

Réalisé par :

EL BAGHDADI MOHAMED

BERRAG AYOUB

Encadré par :

Pr. MEKNASSI MOHAMED

SOMMAIRE

TABLE DES FIGURES

I. INTRODUCTION AU BIG DATA

1. Le phénomène Big Data

L'explosion quantitative des données numériques a obligé les chercheurs à trouver de nouvelles manières de voir et d'analyser le monde. Il s'agit de découvrir de nouveaux ordres de grandeur concernant la capture, la recherche, le partage, le stockage, l'analyse et la présentation des données. Ainsi est né le « Big Data ». Il s'agit d'un concept permettant de stocker un nombre indicible d'informations sur une base numérique. Selon les archives de la bibliothèque numérique de l'Association for Computing Machinery (ou ACM) dans des articles scientifiques concernant les défis technologiques à relever pour visualiser les « grands ensembles de données », cette appellation est apparue en octobre 1997.

2. Définition du Big Data

Littéralement, ces termes signifient mégadonnées, grosses données ou encore données massives. Ils désignent un ensemble très volumineux de données qu'aucun outil classique de gestion de base de données ou de gestion de l'information ne peut vraiment travailler. En effet, nous procréons environ 2,5 trillions d'octets de données tous les jours. Ce sont les informations provenant de partout : messages que nous nous envoyons, vidéos que nous publions, informations climatiques, signaux GPS, enregistrements transactionnels d'achats en ligne et bien d'autres encore. Ces données sont baptisées Big Data ou volumes massifs de données. Les géants du Web, au premier rang desquels Yahoo (mais aussi Facebook et Google), ont été les tous premiers à déployer ce type de technologie.

Cependant, aucune définition précise ou universelle ne peut être donnée au Big Data. Etant un objet complexe polymorphe, sa définition varie selon les communautés qui s'y intéressent en tant qu'usager ou fournisseur de services. Une approche transdisciplinaire permet d'appréhender le comportement des différents acteurs : les concepteurs et fournisseurs d'outils (les informaticiens), les catégories d'utilisateurs (gestionnaires, responsables d'entreprises, décideurs politiques, chercheurs), les acteurs de la santé et les usagers.

Le big data ne dérive pas des règles de toutes les technologies, il est aussi un système technique dual. En effet, il apporte des bénéfices mais peut également générer des inconvénients. Ainsi, il sert aux spéculateurs sur les marchés financiers, de manière autonome avec, à la clé, la constitution des bulles hypothétiques.

L'arrivée du Big Data est maintenant présentée par de nombreux articles comme une nouvelle révolution industrielle semblable à la découverte de la vapeur (début du 19e siècle), de l'électricité (fin du 19e siècle) et de l'informatique (fin du 20e siècle). D'autres, un peu plus mesurés, qualifient ce phénomène comme étant la dernière étape de la troisième révolution industrielle, laquelle est en fait celle de « l'information ». Dans tous les cas, le Big Data est considéré comme une source de bouleversement profond de la société.

3. Big Data : l'analyse de données en masse

Inventé par les géants du web, le Big Data se présente comme une solution dessinée pour permettre à tout le monde d'accéder en temps réel à des bases de données géantes. Il vise à proposer un choix aux solutions classiques de bases de données et d'analyse (plate-forme de Business Intelligence en serveur SQL…).

Selon le Gartner, ce concept regroupe une famille d'outils qui répondent à une triple problématique dite règle des 3V. Il s'agit notamment d'un Volume de données considérable à traiter, une grande Variété d'informations (venant de diverses sources, non-

structurées, organisées, Open…), et un certain niveau de Vélocité à atteindre, autrement dit de fréquence de création, collecte et partage de ces données.

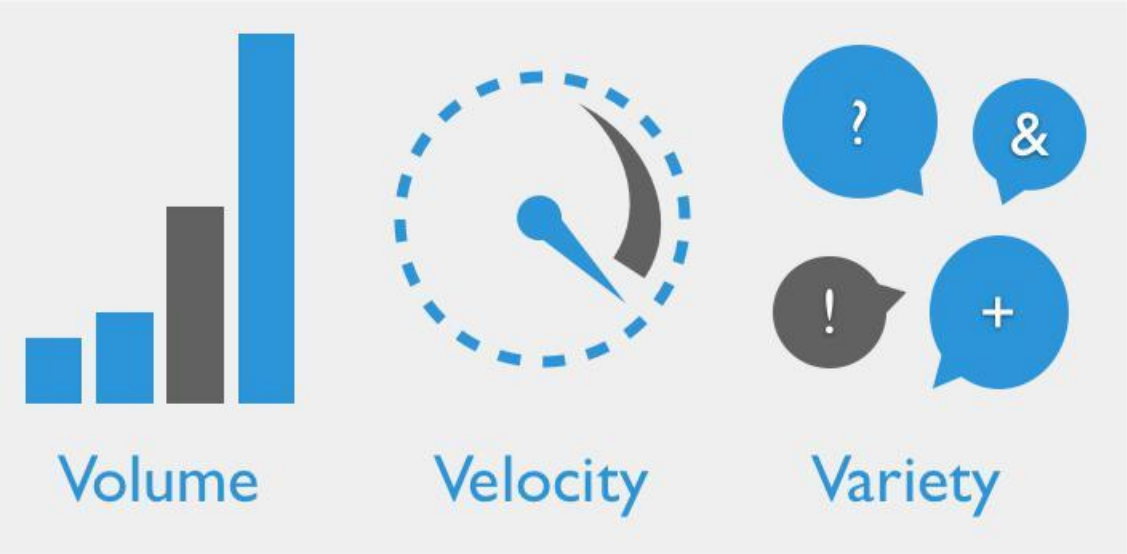

Figure 1 la règle de 3V

4. Les évolutions technologiques derrière le Big Data

Les créations technologiques qui ont facilité la venue et la croissance du Big Data peuvent globalement être catégorisées en deux familles : d'une part, les technologies de stockage, portées particulièrement par le déploiement du Cloud Computing. D'autre part, l'arrivée de technologies de traitement ajustées, spécialement le développement de nouvelles bases de données adaptées aux données non-structurées (Hadoop) et la mise au point de modes de calcul à haute performance (MapReduce).

Il existe plusieurs solutions qui peuvent entrer en jeu pour optimiser les temps de traitement sur des bases de données géantes à savoir les bases de données NoSQL (comme MongoDB, Cassandra ou Redis), les infrastructures du serveur pour la distribution des traitements sur les nœuds et le stockage des données en mémoire :

La première solution permet d'implémenter les systèmes de stockage considérés comme plus performants que le traditionnel SQL pour l'analyse de données en masse (orienté clé/valeur, document, colonne ou graphe).

La deuxième est aussi appelée le traitement massivement parallèle. Le Framework Hadoop en est un exemple. Celui-ci combine le système de fichiers distribué HDFS, la base NoSQL HBase et l'algorithme MapReduce.

Quant à la dernière solution, elle accélère le temps de traitement des requêtes.

5. Evolution du Big Data

Chaque technologie, appartenant au système mégadonnée, a son utilité, ses atouts et ses inconvénients. Etant un milieu en perpétuelle évolution, le Big Data cherche toujours à optimiser les performances des outils. Ainsi, son paysage technologique bouge très vite, et de nouvelles solutions naissent très fréquemment, avec pour but d'optimiser encore plus les technologies existantes. Pour illustrer cette évolution, MapReduce et Spark représentent des exemples très concrets.

Décrit par Google en 2004, MapReduce est un pattern implémenté ultérieurement dans le projet Nutch de Yahoo, qui deviendra le projet Apache Hadoop en 2008. Cet algorithme dispose d'une grande capacité en matière de stockage de données. Le seul hic est qu'il est un peu lent. Cette lenteur est notamment visible sur des volumes modestes. Malgré cela, les solutions, souhaitant proposer des traitements quasi-instantanés sur ces volumes, commencent à

délaisser MapReduce. En 2014, Google a donc annoncé qu'il sera succédé par une solution SaaS dénommée Google Cloud Dataflow.

Spark est aussi une solution emblématique permettant d'écrire simplement des applications distribuées et proposant des bibliothèques de traitement classique. Entre-temps, avec une performance remarquable, il peut travailler sur des données sur disque ou des données chargées en RAM. Certes, il est plus jeune mais il dispose d'une communauté énorme. C'est aussi un des projets Apache ayant une vitesse de développement rapide. En somme, c'est une solution qui s'avère être le successeur de MapReduce, d'autant qu'il a l'avantage de fusionner une grande partie des outils nécessaires dans un cluster Hadoop.

6. Innovations disruptives qui changent la donne

Le Big Data et les analytics sont utilisés dans presque tous les domaines. Ils se sont même construits une place importante dans la société. Ils se traduisent sous plusieurs formes à ne citer que l'usage de statistiques dans le sport de haut niveau, le programme de surveillance PRISM de la NSA, la médecine analytique ou encore les algorithmes de recommandation d'Amazon.

En entreprise particulièrement, l'usage d'outils Big Data & Analytics répond généralement à plusieurs objectifs comme l'amélioration de l'expérience client, l'optimisation des processus et de la performance opérationnelle, le renforcement ou diversification du business model.

De nouvelles opportunités significatives de différenciation concurrentielle sont générées par l'ère de la gestion d'importants volumes de données et de leur analyse. Pour les organisations, plusieurs raisons peuvent les inciter à se tourner vers cette nouvelle administration de données à savoir la gestion rentable des données, l'optimisation du stockage d'informations, la possibilité de faire des analyses programmables ou encore la facilité de la manipulation des données.

7. Big Data dans les fonctions Marketing et commerciales

Cette technologie représente aux yeux de tous un enjeu commercial privilégié compte tenu de sa capacité à impacter le commerce en profondeur dans l'économie mondiale intégrée. En effet, les entreprises, peu importe leur taille, font partie des premières à bénéficier des avantages obtenus à partir d'une bonne manipulation des données massives.

Cependant, les mégadonnées jouent également un rôle essentiel dans la transformation des processus, de la chaîne logistique, des échanges de type « Machine-to-Machine » dans le but de développer un meilleur « écosystème informationnel ». Ils permettent aussi de prendre des décisions plus véloces et plus crédibles, prenant en considération des informations internes mais également externes
à l'organisation. Ils peuvent entre-temps servir d'appui pour la
gestion des risques et de la fraude.

8. Comment trier le bon grain de l'ivraie ?

Comme le dit le vieil adage « trop d'informations tuent l'information ». Il s'agit en fait du principal problème avec les mégadonnées. La quantité énorme des informations est un des obstacles. L'autre obstacle provient évidemment du niveau de certitude qu'on peut avoir sur une donnée.

En effet, les données qui découlent du marketing numérique peuvent être considérées comme des informations « incertaines », dans la mesure par exemple où on ne peut être sûr de l'identité de qui est en train de cliquer sur une offre incluse dans une URL. Le volume de données associé au manque de crédibilité de celles-ci rend son exploitation plus alambiquée.

Pour autant, grâce aux algorithmes statistiques, des solutions existent. C'est d'ailleurs, avant même de se demander s'il serait possible de collecter et stocker le big data, qu'on devrait toujours commencer par s'interroger de son aptitude à les analyser et de leur utilité.

Avec un but convenablement déterminé et des données d'une qualité suffisante, les algorithmes et méthodes statistiques permettent désormais de concevoir de la valeur alors que ce n'était pas encore faisable il y a encore quelques années. A ce propos, on peut distinguer deux types d'écoles dans le domaine prédictif à savoir l'intelligence artificielle ou « machine learning » et la statistique. Ces deux secteurs bien qu'ils soient distincts se rejoignent finalement de plus en plus. De plus, ils peuvent être utilisés en simultanéité de manière vertueuse et intelligente pour mener à bien un projet.

Là où l'usage des mégadonnées en gestion devient un enjeu vital pour les entreprises.

Parmi les utilisateurs les plus enthousiastes du Big Data, on retrouve les gestionnaires et les économistes. Ces derniers définissent ce phénomène par la règle des 5V (Volume, Velocity, Variety, Veracity, Value).

Figure 2 les 5V de Big Data

Le volume

Le volume correspond à la masse d'informations produite chaque seconde. Selon des études, pour avoir une idée de l'accroissement exponentiel de la masse de données, on considère que 90 % des données ont été engendrées durant les années où l'usage d'internet et des réseaux sociaux a connu une forte croissance. L'ensemble de toutes les données produites depuis le début des temps jusqu'à la fin de l'année 2008, conviendrait maintenant à la masse de celles qui sont générées chaque minute. Dans le monde des affaires, le volume de données collecté chaque jour est d'une importance vitale.

La vélocité

La vélocité équivaut à la rapidité de l'élaboration et du déploiement des nouvelles données. Par exemple, si on diffuse des messages sur les réseaux sociaux, ils peuvent devenir « viraux » et se répandre en un rien de temps. Il s'agit d'analyser les données au décours de leur lignée (appelé parfois analyse en mémoire) sans qu'il soit

indispensable que ces informations soient entreposées dans une base de données.

La variété

Seulement 20% des données sont structurées puis stockées dans des tables de bases de données relationnelles similaire à celles utilisées en gestion comptabilisée. Les 80% qui restent sont non-structurées. Cela peut être des images, des vidéos, des textes, des voix, et bien d'autres encore… La technologie Big Data, permet de faire l'analyse, la comparaison, la reconnaissance, le classement des données de différents types comme des conversations ou messages sur les réseaux sociaux, des photos sur différents sites etc. Ce sont les différents éléments qui constituent la variété offerte par le Big Data.

La véracité

La véracité concerne la fiabilité et la crédibilité des informations collectées. Comme le Big Data permet de collecter un nombre indéfini et plusieurs formes de données, il est difficile de justifier l'authenticité des contenus, si l'on considère les post Twitter avec les abréviations, le langage familier, les hashtags, les coquilles etc. Toutefois, les génies de l'informatique sont en train de développer de nouvelles techniques qui devront permettre de faciliter la gestion de ce type de données notamment par le W3C.

La valeur

La notion de valeur correspond au profit qu'on puisse tirer de l'usage du Big Data. Ce sont généralement les entreprises qui commencent à obtenir des avantages incroyables de leurs Big Data. Selon les gestionnaires et les économistes, les entreprises qui ne s'intéressent pas sérieusement au Big Data risquent d'être pénalisées et écartées.

Puisque l'outil existe, ne pas s'en servir conduirait à perdre un privilège concurrentiel.

9. L'essor des mégadonnées en médecine

La médecine est un art qui use des sciences. En effet, un médecin praticien est en simultanéité un scientifique qui a obtenu des connaissances en biophysique, sémiologie médicale et chirurgicale, anatomie, biochimie, physiologie, biologie, … et un artiste qui maîtrise des habiletés pour effectuer des gestes thérapeutiques adaptés. Désormais, les connaissances traditionnelles ne suffisent plus pour mieux amplifier le pouvoir d'un médecin dans l'investigation et le soin. Il a également appris à maîtriser des technologies de plus en plus sophistiquées dans les différentes spécialités médicales. On assiste à l'essor du génie biologique médical ou GBM. Cette alternative offre aux médecins de nouvelles possibilités de diagnostic à savoir, les appareils d'imagerie : scintigraphie, échographes, imagerie par résonance magnétique (IRM) etc. Les automates d'analyse biologique, les appareils d'analyse de signaux comme l'électrocardiogramme (ECG) ou encore l'électroencéphalogramme (EEG), ainsi que les appareils de traitement des pathologies (dialyse, laser, assistance respiratoire, médecine nucléaire…) figurent aussi parmi les fruits de l'alliance technologie/médecine.

Majoritairement pilotés par des ordinateurs spécialisés qui sont directement ou indirectement connectés à un réseau informatique, ces dispositifs permettent de collecter des informations diverses concernant les patients. Ils se présentent comme de nouveaux moyens d'investigation, d'acquisition et de stockage de données, de comparaison de l'information que les médecins traitants peuvent mettre en œuvre afin d'accroître leur réactivité dans les différentes étapes cliniques essentielles à la prise en charge de leurs patients. Ils peuvent aussi s'en servir pour mener des études épidémiologiques des maladies dans la population.

10. L'avenir du Big Data

Etant une tendance lourde, le Big Data n'est pas une mode. Dans le domaine de l'usage, il satisfait une nécessité de travailler la donnée plus profondément, pour créer de la valeur, conjointement à des aptitudes technologiques qui n'existaient pas dans le passé. Cependant, compte tenu de l'évolution des technologies qui ne semble pas vouloir s'estomper, on ne peut pas alors parler d'une norme véritable ou de standards dans le domaine du Big data.

Beaucoup d'applications du Big Data n'en sont qu'à leurs préludes et on peut s'attendre à voir apparaître des utilisations auxquelles on ne s'attend pas encore aujourd'hui. En quelque sorte, le Big Data est un tournant pour les organisations au moins aussi important qu'internet en son temps. Chaque entreprise doit donc s'y mettre dès maintenant. Dans le cas contraire, il y a un risque qu'elle se rendent comptent d'ici quelques années qu'elles se sont faites dépasser par la concurrence. Les gouvernements et les organismes publics se penchent également sur la question à travers l'open data.

11. Les données massives

D'ici quelques années, le marché du big data va se mesurer en centaines de milliards de dollars. C'est un nouvel eldorado pour le business. Selon des études, il s'agit même d'une vague de fond où l'on retrouve la combinaison de la BI (business intelligence), de l'analytics et de l'internet des objets. IDC affirme qu'il devrait passer au-delà des 125 milliards de dollars avant la fin 2015. En effet, plusieurs études affluent sur cette affirmation et toutes confirment que les budgets que les entreprises vont consacrer au Big Data ne vont connaître que des fortes progressions. Ainsi, rien que le marché des solutions visuelles de découvertes des informations liées à la gestion des données massives va grimper de 2,5 fois plus rapidement que celui des solutions de BI d'ici à 2018.

D'après le calcul effectué par le cabinet Vanson Bourne, dans le monde, l'ensemble des dépenses consacrées au Big data, dans les budgets IT des grandes entreprises, devrait représenter un quart du budget total IT en 2018, s'il en est encore à 18% actuellement. Le Cap Gemini a aussi commandité une étude en mars 2015. Le résultat a montré que 61% des entreprises sont conscientes de l'utilité du Big Data en tant que « moteur de croissance à part entière ». De ce fait, on lui accorde beaucoup plus d'importance que leurs produits et services existants. Cette même étude a encore indiqué que 43% d'entre elles se sont déjà réorganisées ou se restructurent présentement pour exploiter le potentiel du Big Data.

II. LES ALGORITHMES DE BIG DATA LES PLUS UTILISES

Pour profiter des bienfaits de l'analyse de données, il convient de connaître les algorithmes à utiliser. Voici les algorithmes les plus utilisés pour le Big Data Analytics.

Le Big Data peut-être très utile pour les entreprises. La plupart des organisations multinationales s'en remettent de nos jours à l'analyse de données pour aiguiller leurs décisions et ainsi stimuler leur croissance, augmenter leur chiffre d'affaires ou découvrir de nouvelles opportunités.

Toutefois, beaucoup d'entreprises qui souhaitent à leur tour exploiter le Big Data ne savent pas par où commencer. Si tel est votre cas, voici les algorithmes Big Data les plus utilisés.

1. Régression linéaire

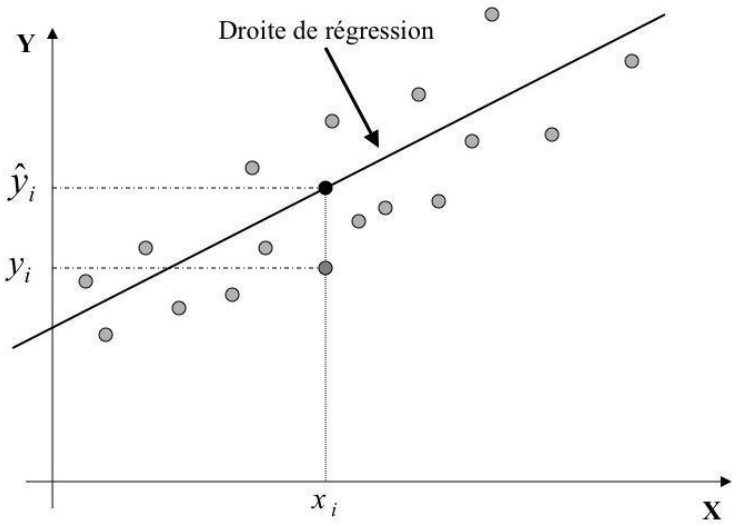

Figure 3 régression linéaire

La régression linéaire est l'algorithme le plus basique et l'un des plus utilisés dans le domaine de l'analyse de données et du Machine Learning. Cet algorithme utilise la relation entre deux ensembles de mesures quantitatives continues.

Le premier ensemble est appelé » prédicteur » ou » variable indépendante «. Le second est appelé » réponse » ou » variable dépendante «. L'objectif de la régression linéaire est d'identifier la relation entre ces deux ensembles sous la forme d'une formule. Une fois la relation quantifiée, la variable dépendante peut être prédite pour n'importe quelle instance de la variable indépendante.

2. Régression logistique

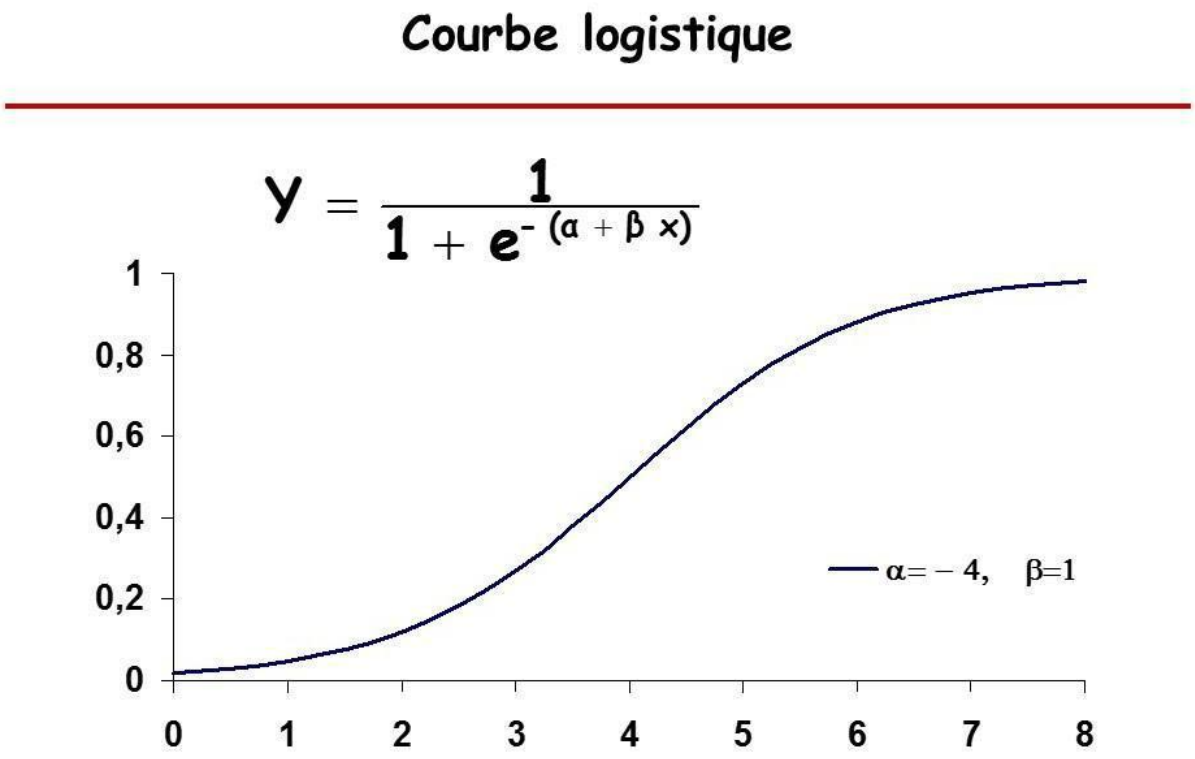

Figure 4 régression logistique

La régression logistique est similaire à la régression linéaire, mais elle est utilisée pour des problèmes de catégorisation plutôt que pour des

prédictions quantitatives. L'objectif de cet algorithme est de déterminer si une instance d'une variable entre ou non dans une catégorie.

Ainsi, le résultat d'une régression logistique est une valeur comprise entre 0 et 1. Plus le résultat est proche de 1, plus la variable entre dans la catégorie. Au contraire, un résultat proche de 0 indique une probabilité que la variable n'entre pas dans la catégorie.

3. Arbres de classification et de régression

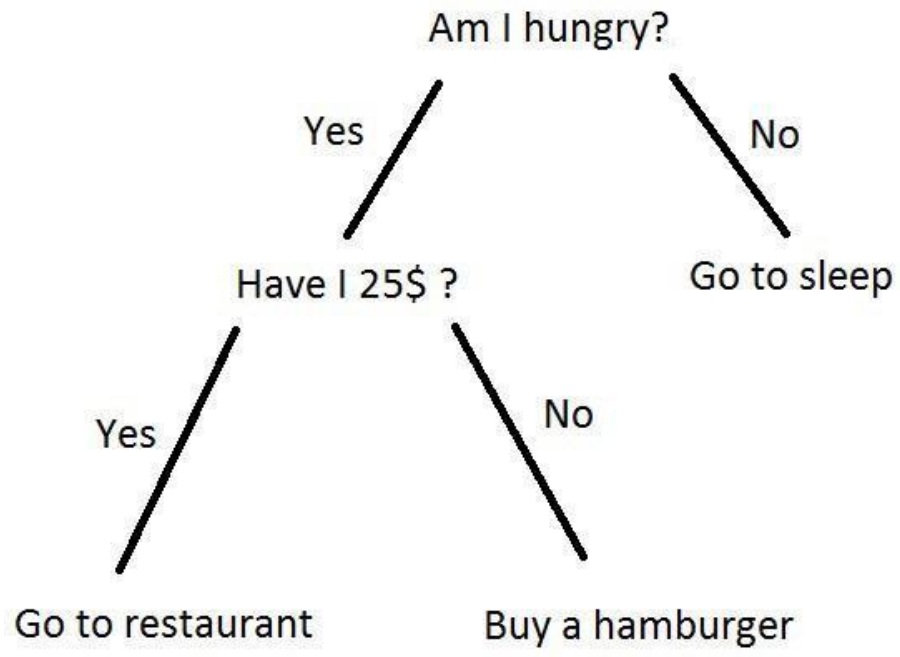

Figure 5 arbre de décision

Les arbres de classification et de régression utilisent une décision pour catégoriser les données. Chaque décision est basée sur une question liée à l'une des variables entrantes. En fonction des réponses, l'instance de donnée est catégorisée. Cette succession de questions

et de réponses et les divisions qui en découlent créent une structure en forme d'arbre.

Bien entendu, les arbres de classification peuvent vite devenir très larges et complexes. L'une des méthodes permettant de contrôler cette complexité est de supprimer certaines questions. Une variante des arbres de classification et de régression est celle des forêts aléatoires. Elle consiste à créer un cumul de petits arbres simples plutôt qu'un seul arbre avec de nombreuses branches. Chacun de ces petits arbres évalue une partie des données, puis les résultats sont assemblés pour créer une prédiction finale.

4. Méthode des k plus proches voisins

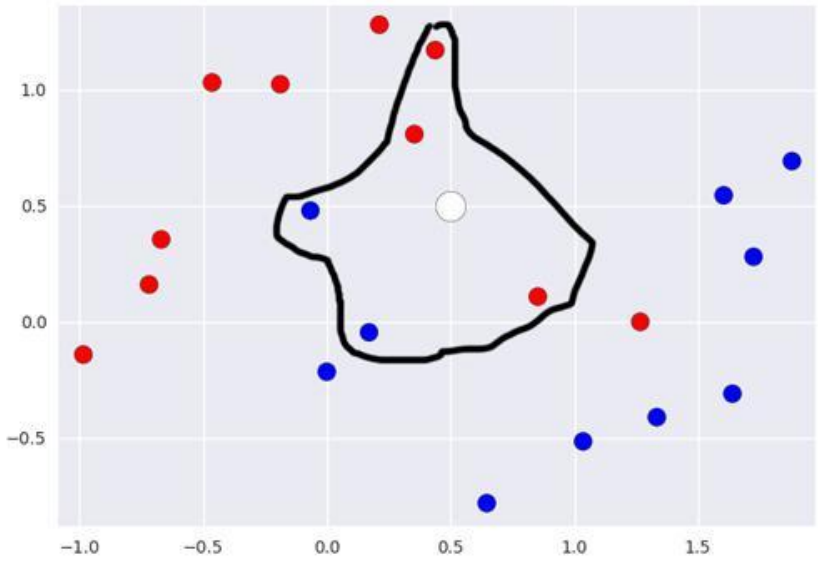

Figure 6 k plus proches voisins

La méthode des k plus proches voisins est également un algorithme de classification. Dans un premier temps, on utilise un ensemble de données pour entraîner l'algorithme. Par la suite, la distance qui

sépare les données d'entraînement des nouvelles données est évaluée pour catégoriser les nouvelles données.

En fonction de la taille de l'ensemble d'entraînement, cet algorithme peut nécessiter beaucoup de ressources de calcul. Il est toutefois souvent utilisé pour sa simplicité d'usage, sa facilité d'entrainement, et la facilité à interpréter les résultats.

5. Partitionnement en K-moyennes

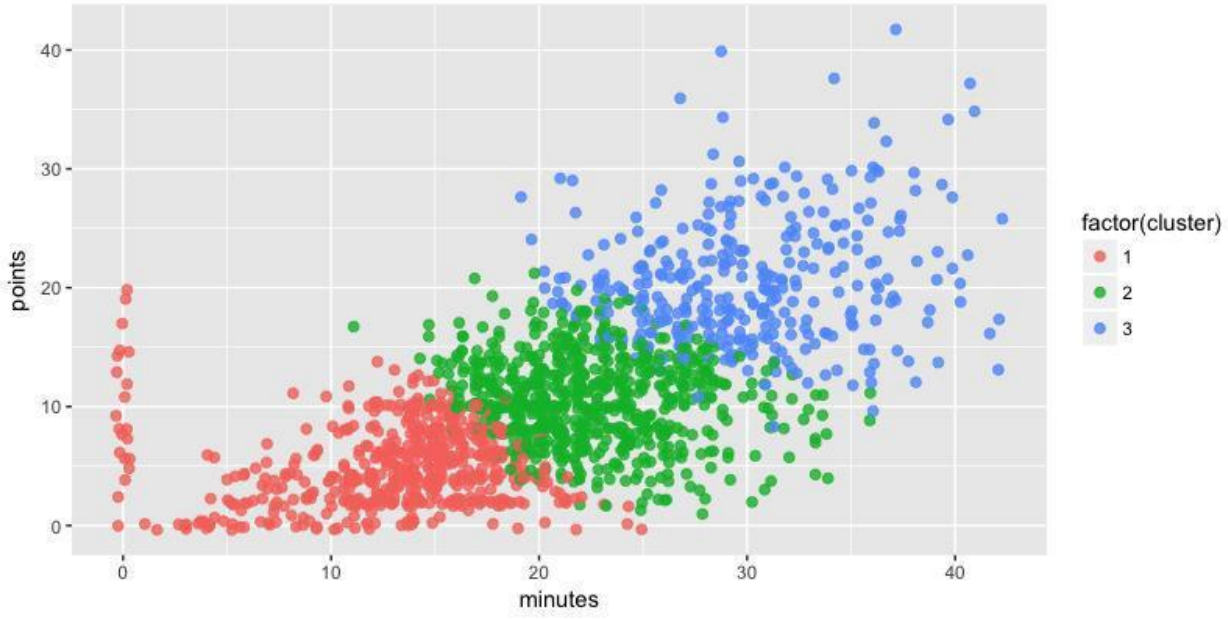

Figure 7 partitionnement en k-moyennes

Le partitionnement en K-moyennes est une méthode qui consiste à créer des groupes d'attributs relatifs. Ces groupes sont appelés partitions. Une fois qu'ils sont créés, les autres instances peuvent être évaluées par rapport à eux afin de déterminer quelle catégorie leur correspond le mieux.

Cette technique est souvent utilisée dans le cadre de l'exploration de données. Les analystes définissent d'abord le nombre de partitions,

puis les données sont réparties en fonction de leurs similitudes. Les partitions diffèrent des catégories, car il s'agit seulement d'instances liées de variables entrantes. Une fois identifiées et analysées, les partitions peuvent toutefois être converties en catégories. Cette méthode est souvent utilisée pour sa simplicité et sa vitesse.

Chacun de ces algorithmes a ses avantages et ses inconvénients, et il est important de choisir le plus adapté en fonction de chaque situation.

III. MACHINE LEARNING ET BIG DATA

1. Définition et explications

Le Machine Learning est une technologie d'intelligence artificielle permettant aux ordinateurs d'apprendre sans avoir été programmés explicitement à cet effet. Pour apprendre et se développer, les ordinateurs ont toutefois besoin de données à analyser et sur lesquelles s'entraîner. De fait, le Big Data est l'essence du Machine Learning, et c'est la technologie qui permet d'exploiter pleinement le potentiel du Big Data. Découvrez pourquoi cette technique et le Big Data sont interdépendants.

2. Le Machine Learning

Si le Machine Learning ne date pas d'hier, sa définition précise demeure encore confuse pour de nombreuses personnes. Concrètement, il s'agit d'une science moderne permettant de découvrir des patterns et d'effectuer des prédictions à partir de données en se basant sur des statistiques, sur du forage de données, sur la reconnaissance de patterns et sur les analyses prédictives. Les premiers algorithmes sont créés à la fin des années 1950. Le plus connu d'entre eux n'est autre que le Perceptron.

Le Machine Learning est très efficace dans les situations où les insights doivent être découvertes à partir de larges ensembles de données diverses et changeantes, c'est à dire : le Big Data. Pour l'analyse de telles données, il se révèle nettement plus efficace que les méthodes traditionnelles en termes de précision et de vitesse. Par exemple, pour en se basant sur les informations associées à une transaction comme le montant et la localisation, et sur les données historiques et sociales, le Machine Learning permet de détecter une fraude potentielle en une milliseconde. Ainsi, cette méthode est nettement plus efficace que les méthodes traditionnelles pour l'analyse de données

transactionnelles, de données issues des réseaux sociaux ou de plateformes CRM.

3. L'utilisation de Machine Learning avec Big Data

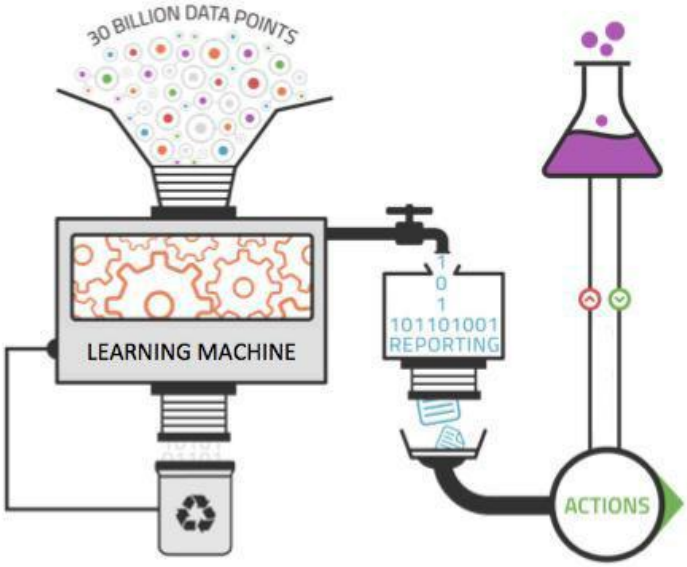

Figure 8 big data et machine learning

Les outils analytiques traditionnels ne sont pas suffisamment performants pour exploiter pleinement la valeur du Big Data. Le volume de données est trop large pour des analyses compréhensives, et les corrélations et relations entre ces données sont trop importantes pour que les analystes puissent tester toutes les hypothèses afin de dégager une valeur de ces données.

Les méthodes analytiques basiques sont utilisées par les outils de business intelligence et de reporting pour le rapport des sommes, pour faire les comptes et pour effectuer des requêtes SQL. Les traitements analytiques en ligne sont une extension systématisée de ces outils analytiques basiques qui nécessitent l'intervention d'un humain pour spécifier ce qui doit être calculé.

Le Machine Learning est idéal pour exploiter les opportunités cachées du Big Data. Cette technologie permet d'extraire de la valeur en provenance de sources de données massives et variées sans avoir besoin de compter sur un humain. Elle est dirigée par les données, et convient à la complexité des immenses sources de données du Big Data. Contrairement aux outils analytiques traditionnels, il peut également être appliqué aux ensembles de données croissants. Plus les données injectées à un système Machine Learning sont nombreuses, plus ce système peut apprendre et appliquer les résultats
à des insights de qualité supérieure. Le Machine Learning permet ainsi de découvrir les patterns enfouis dans les données avec plus d'efficacité que l'intelligence humaine.

4. Pourquoi le Machine Learning n'est rien sans Big Data

Sans le Big Data, le Machine Learning et l'intelligence artificielle ne seraient rien. Les données sont l'instrument qui permet à l'IA de comprendre et d'apprendre à la manière dont les humains pensent. C'est le Big Data qui permet d'accélérer la courbe d'apprentissage et permet l'automatisation des analyses de données. Plus un système Machine Learning reçoit de données, plus il apprend et plus il devient précis.

L'intelligence artificielle est désormais capable d'apprendre sans l'aide d'un humain. Par exemple, l'algorithme Google DeepMind a récemment appris seul à jouer à 49 jeux vidéo Atari. Par le passé, le développement était limité par le manque d'ensembles de données disponibles, et par son incapacité à analyser des quantités massives de données en quelques secondes.

Aujourd'hui, des données sont accessibles en temps réel à tout moment. Ceci permet à l'IA et au Machine Learning de passer à une approche dirigée par les données. La technologie est désormais suffisamment agile pour accéder aux ensembles de données colossaux et pour les analyser. De fait, des entreprises de toutes les

industries se joignent désormais à Google et Amazon pour implémenter des solutions IA pour leurs entreprises.

Un exemple de Machine learning appliqué ? MetLife, l'un des principaux assureurs d'entreprise à l'échelle mondiale, utilise cette technique et le Big Data pour optimiser son activité. La reconnaissance de discours lui a permis d'améliorer le tracking d'accidents et de mieux mesurer leurs conséquences. Le traitement de réclamations est désormais mieux pris en charge car les modèles de réclamations ont été enrichis à l'aide de données non structurées qui peuvent être analysées par le biais de cette technologie.

Autre exemple, cette technique permet d'apprendre les habitudes des occupants d'un foyer. Les concepteurs d'objets connectés, notamment de thermostats, peuvent analyser la température du logement afin de comprendre la présence et l'absence des occupants pour couper le chauffage et le rallumer quelques minutes avant leur retour.

5. Le Deep Learning

L'apprentissage automatique est un sous domaine de l'intelligence artificielle. Le Deep Learning est lui-même une sous-catégorie de l'apprentissage automatique. L'exemple d'application le plus commun est la reconnaissance visuelle. Par exemple, un algorithme va être programmer pour détecter certains visages depuis les images en provenance d'une caméra. Suivant la base de données attribuée, il pourra repérer un individu recherché dans une foule, détecter le taux de satisfaction à la sortie d'un magasin en détectant les sourires, etc. Un ensemble d'algorithme pourra également reconnaître la voix, le ton, l'expression d'un questionnement, d'une affirmation et les mots.

Pour ce faire, le Deep Learning repose principalement sur la reproduction d'un réseau neuronal inspiré des systèmes cérébraux présents dans la nature. Les développeurs décident suivant

l'application souhaité quel type d'apprentissage ils vont mettre en place. Dans ce cadre, on parle d'apprentissage supervisé, d'apprentissage non supervisé dans lequel la machine va se nourrir de données non sélectionnées au préalable, semi-supervisé, par renforcement (lié à une observation), ou par transfert dans laquelle les algorithmes vont appliquer une solution apprise dans une situation jamais vue.

En revanche, cette technique a besoin de beaucoup de données pour s'entraîner et obtenir des taux de réussite suffisant pour être utiliser. Un Lac de données ou Data Lake est essentiel pour parfaire l'apprentissage des algorithmes de Deep Learning. L'apprentissage profond nécessite également une puissance de calcul supérieure pour réaliser son office.

6. Les analyses prédictives donnent du sens au Big Data

Les analyses prédictives consistent à utiliser les données, les algorithmes statistiques et les techniques de Machine Learning pour prédire les probabilités de tendances et de résultats financiers des entreprises, en se basant sur le passé. Elles rassemblent plusieurs technologies et disciplines comme les analyses statistiques, le data mining, le modelling prédictif et le Machine Learning pour prédire le futur des entreprises. Par exemple, il est possible d'anticiper les conséquences d'une décision ou les réactions des consommateurs.

Les analyses prédictives permettent de produire des insights exploitables à partir de larges ensembles de données, pour permettre aux entreprises de décider quelle direction emprunter par la suite et offrir une meilleure expérience aux clients. Grâce à l'augmentation du nombre de données, de la puissance informatique, et du développement de logiciels IA et d'outils analytiques plus simples à utiliser, comme Salesforce Einstein, un grand nombre d'entreprises peuvent désormais utiliser les analyses prédictives.

Selon une étude menée par Bluewolf auprès de 1700 clients de Salesforce, 75% des entreprises qui augmentent leurs investissements dans les technologies analytiques en tirent profit. 81% de ces utilisateurs des produits Salesforce estiment que l'utilisation des analyses prédictives est l'initiative la plus importante de leur stratégie de ventes. Les analyses prédictives permettent d'automatiser les prises de décision, et donc d'augmenter la rentabilité et la productivité d'une entreprise.

L'intelligence artificielle et le Machine Learning représentent le niveau supérieur des analyses de données. Les systèmes informatiques cognitifs apprennent constamment sur l'entreprise et prédisent intelligemment les tendances de l'industrie, les besoins des consommateurs et bien plus encore. Peu d'entreprises ont déjà atteint le niveau des applications cognitives, défini par quatre caractéristiques principales : la compréhension des données non structurées, la possibilité de raisonner et d'extraire des idées, la capacité à affiner l'expertise à chaque interaction, et la capacité à voir, parler et entendre pour interagir avec les humains de façon naturelle. Pour cela, il convient de développer le traitement par algorithme des langages naturels.

7. L'apprentissage automatique au service du Data Management

Face à l'augmentation massive du volume de données stockées par les entreprises, ces dernières doivent faire face à de nouveaux défis. Parmi les principaux challenges liés au Big Data, on dénombre la compréhension du Dark Data, la rétention de données, l'intégration de données pour de meilleurs résultats analytiques, et l'accessibilité aux données. Le Machine Learning peut s'avérer très utile pour relever ces différents défis.

Toutes les entreprises accumulent au fil du temps de grandes quantités de données qui demeurent inutilisées. Il s'agit des dark data.

Grâce au Machine Learning et aux différents algorithmes, il est possible de faire le tri parmi ces différents types de données stockées sur les serveurs. Par la suite, un humain qualifié peut passer en revue le schéma de classification suggéré par l'intelligence artificielle, y apporter les changements nécessaires, et le mettre en place.

Pour la rétention de données, cette pratique peut également s'avérer efficace. L'intelligence artificielle peut identifier les données qui ne sont pas utilisées et suggérer lesquelles peuvent être supprimées. Même si les algorithmes n'ont pas la même capacité de discernement que les êtres humains, le Machine Learning permet de faire un premier tri dans les données. Ainsi, les employés économisent un temps précieux avant de procéder à la suppression définitive des données obsolètes.

Cette technologie est aussi utile pour l'intégration de données. Pour tenter de déterminer le type de données qu'ils doivent agréger pour leurs requêtes, les analystes créent généralement un répertoire dans lequel ils placent différents types de données en provenance de sources variées pour créer un bassin de données analytique. Pour ce faire, il est nécessaire de développer des méthodes d'intégration pour accéder aux différentes sources de données en provenance desquelles ils extraient les données. Cette technique peut faciliter le processus en créant des mappings entre les sources de données et le répertoire. Ceci permet de réduire le temps d'intégration et d'agrégation.

Enfin, l'apprentissage des données permet d'organiser le stockage de données pour un meilleur accès. Au cours des cinq dernières années, les vendeurs de solutions de stockage de données ont mis leurs efforts dans l'automatisation de la gestion de stockage. Grâce à la réduction de prix du SSD, ces avancées technologiques permettent aux départements informatiques d'utiliser des moteurs de stockage intelligents reposant sur le machine Learning pour voir quels types de données sont utilisés le plus souvent et lesquels ne sont pratiquement jamais utilisé. L'automatisation peut être utilisée pour

stocker les données en fonction des algorithmes. Ainsi, l'optimisation n'a pas besoin d'être effectuée manuellement.

Figure 9 data management

IV. LES APPLICATIONS DU MACHINE LEARNING EN BIG DATA

La rentabilité du Big Data réside en grande partie, dans la capacité de l'entreprise à analyser ses données afin d'en tirer des informations utiles. Alors que 2012 a été l'année d'avènement du Big Data, depuis 2013 on parle de « Big Data Analytics ». Volume, Vitesse et Variété, voilà ce qui caractérise les données collectées par les entreprises aujourd'hui. A la question de comment les analyser, la réponse est le « Machine Learning ».

1. Les spams de nos boîtes mails

Lorsque nous déplaçons un mail dans les spams, la boîte mail retient en mémoire les caractéristiques de ce mail. Grâce aux algorithmes de « Machine Learning », lorsqu'un mail présentant des caractéristiques similaires nous parviendra à nouveau, notre boîte mail le mettra directement dans nos spams. La boîte mail fait office de garde du corps personnalisé face aux mails que nous jugeons inopportuns.

2. Le e-commerce et l'exemple Amazon

Amazon le site de vente est un exemple intéressant d'application du « Machine Learning » dans l'e-commerce. Supposons par exemple que nous effectuons la recherche d'un produit sur Amazon aujourd'hui.

Lorsque nous revenons un autre jour sur le site, il est capable de nous proposer des produits en rapport avec nos besoins spécifiques. Et ceci grâce à des algorithmes de « Machine Learning » qui prévoient l'évolution de nos besoins à partir de nos précédentes visites sur le site.

3. IBM et la prise de notes dans le domaine médical

IBM a réussi à extraire à partir des notes prises par des médecins pendant les consultations (notes électroniques), des critères pour diagnostiquer l'insuffisance cardiaque. Ils ont développé un algorithme de « Machine Learning » qui synthétise le texte en utilisant une technique appelée « Natural Language Processing » (NLP).

De la même manière qu'un cardiologue peut lire les notes d'un autre médecin et déterminer si un patient a une insuffisance cardiaque, les ordinateurs peuvent maintenant faire la même chose.

4. Une application pour identifier les facteurs déclencheurs de migraines

Healint une start-up à Singapour, a lancé sur Android l'application mobile Migraine Buddy. Cette application permet aux patients migraineux d'identifier les facteurs déclencheurs des migraines et les différents symptômes. L'application utilise des algorithmes de «Machine Learning» qui permettent de mettre en évidence des corrélations entre habitudes de vie et migraines.

Le système permet également de comparer l'efficacité des différents moyens pour soulager les douleurs. Les rapports dynamiques que l'application génère permettent d'évaluer au fil des mois le contrôle que nous exerçons sur les migraines, et de transmettre aux médecins des données permettant de l'aider à ajuster son traitement.

Le « Machine Learning » consiste donc à apprendre, en tirant des prévisions de fonctionnement ou de comportement à partir des données.

Le Big Data et le « Machine Learning » représentent ainsi la prochaine révolution informatique. Et comme chaque révolution, cette dernière va transformer notre manière de vivre, de travailler et de penser.

REFERENCES

https://www.lebigdata.fr/

https://bigdata-madesimple.com/

https://www.ee.columbia.edu/~cylin/course/bigdata/EECS6893-BigDataAnalytics-Lecture4.pdf

https://www.futura-sciences.com/tech/

https://fr.coursera.org/

https://www.lemagit.fr/