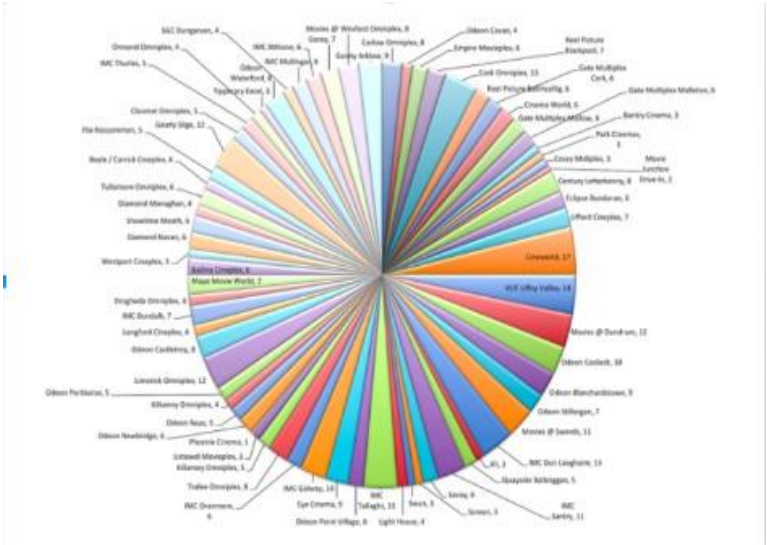
Big Data Analytics

Présenté par

Munaf Abbas

Qu’est-ce que les Big Data ?

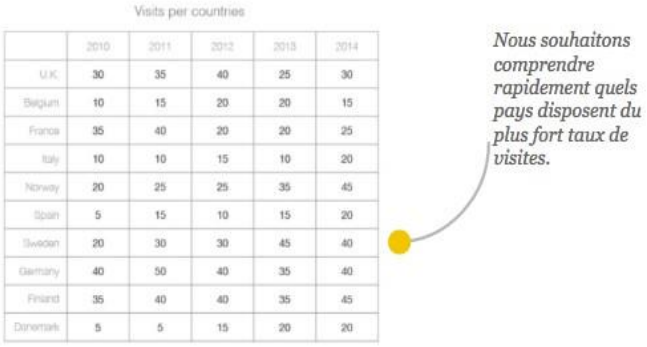
Le Big Data est un phénomène qui a vu le jour avec l’abondance des informations provenant des nouvelles technologies d’information et de communication. Appelées communément « données », ces informations variées nécessitent désormais une nouvelle approche pour la compréhension, le traitement, le positionnement et le stockage. Avec environ 21 milliards d’objets connectés en 2020[1], les Big Data atteignent un niveau de complexité sans précédent. D’énormes quantités de données proviennent à titre d’exemple des tweets stockés sur le serveur de Twitter, des informations de géolocalisation fournies par Google, des données d’assurance maladie, des types et les lieux d’achats repérés sur les cartes de crédits, des différentes informations concernant les abonnements de streaming, de la gestion de l’énergie dans les bâtiments, du domaine de l’industrie, de l’agriculture, de la sécurité ou encore du secteur du transport.



Les Big Data Analytics se sont ainsi imposées pour contribuer à la compréhension de ces flux permanents de données, en générant de matrices de comportements appropriés aux besoins et aux différents domaines d’application. Cette analyse est généralement produite autour de deux axes majeurs : les éléments du réseau (hommes ou machines) et l’appropriation de la solution par rapport à l’utilisateur final : soit des applications de « grand-public » issues des réseaux

[1] https://www.lebigdata.fr/iot-big-data

hommes-hommes ou hommes-machines, soit des applications en matière d'industrie ou de politiques publiques issues des réseaux machines-machines ou hommes-machines. S'ajoute au grand volume et à la variété des mégadonnées une troisième caractéristique inhérente des Big Data à savoir, la vélocité se référant à la vitesse de traitement de ces données.



Le Fast Data désigne l'application des analyses Big Data à de plus petits ensembles de données, en temps réel ou presque réel. Pour procéder aux analyses Fast Data, plusieurs conditions sont nécessaires. Il est tout d'abord indispensable de disposer d'un système de streaming capable de transférer les données dès qu'elles sont générées. En outre, il est nécessaire de disposer d'une Data Warehouse capable de traiter et d'analyser les données qu'elles y sont stockées.
Les caméras de surveillance connectées, qui enregistrent des événements en continu, peuvent aussi utiliser le Fast Data. L'analyse de données en temps réel leur permet d'identifier instantanément les anomalies de sécurité. En somme, le Fast Data s'avère utile pour toutes les situations où les données doivent être analysées dès lors qu'elles sont générées.

Contexte et enjeux actuels

Des nouvelles approches ont favorisé les mises en application du Big Data. Nous parlons aujourd'hui d'Internet Of Things (IoT), de Machine Learning (ML) et d'Artificial Intelligence (AI).

L'Internet of Things dépasse le stade de l'Internet Of People, où l'interaction se faisait entre les utilisateurs via la machine, pour entrer dans une époque où les appareils communiquent directement les uns avec les autres sans intervention humaine. Le même appareil peut aujourd'hui surveiller le temps qu'il fait, accéder à la température, envoyer un message aux personnes concernées et prendre la décision d'ajuster la température des locaux. Plus les appareils fonctionnent de cette manière, plus il y a de données générées. L'Internet des Objets concerne les objets sont capables d'échanger des informations et de communiquer entre eux. Des objets capables aussi de communiquer et d'interagir avec leurs utilisateurs en utilisant Internet mais aussi d'autres réseaux de communication bien moins connus mais tous efficaces.

Le Machine Learning fait référence à la capacité d'un ordinateur d'apprendre à partir de données. L'apprentissage machine est à la base des recommandations de contenu sur YouTube et Netflix, de la recherche de Google ou du fil de nouvelles de Facebook. Ce ne sont là que quelques exemples de la façon dont les algorithmes d'apprentissage machine influencent nos expériences quotidiennes.

L'intelligence artificielle est l'étape suivante après l'apprentissage machine. Ici, non seulement un ordinateur apprend à partir de données, mais il utilise cette information pour prendre ses propres décisions et façonner son propre comportement. Microsoft et Google ont tous deux montré leurs efforts pour créer des robots humanoïdes. Facebook utilise l'intelligence artificielle pour aider à prévenir les suicides. La technologie progresse à un rythme où il y a eu plusieurs cas où la pensée d'un ordinateur a surpassé celle d'un humain.

La Chine a annoncé vouloir investir 150 milliards de dollars dans l'intelligence artificielle (IA) d'ici à 2030, les États-Unis ont répondu un an et demi après qu'ils feraient du développement de l'IA une priorité nationale et que le financement serait à la hauteur de cette ambition[2].

De son côté, l'Europe semble loin derrière avec seulement 20 milliards d'euros d'investissements prévus à l'horizon 2020. Pour beaucoup, l'avenir de l'intelligence artificielle serait donc en train

[2] https://www.lesechos.fr/idees-debats/cercle/intelligence-artificielle-pourquoi-leurope-a-sa-carte-a-jouer-1124917

de se jouer entre les États-Unis et la Chine. Mais la réalité apparaît plus complexe. Surtout, avec sa politique d'IA responsable, l'Europe n'a pas dit son dernier mot.

Domaines d'application des Big Data

Les avantages des Big Data se traduisent par les multiples secteurs d'usages en question. Nous abordons dans ce qui suit, certains contextes où les Big Data, mises en application, aident à l'optimisation de certains secteurs en termes de résultats et de réalisations.

L'industrie des soins de santé connait depuis quelques années une réelle progression au niveau de l'intégration, malgré les supports hétérogènes des informations. Les assureurs et les fournisseurs s'efforcent de combiner des données provenant de différentes sources, comme les demandes de règlement, les radiographies, les notes du médecin et les ordonnances.

L'industrie de la finance quant à elle, est tout-en-un sur l'idée de prendre des décisions basées sur l'analyse informatique. Les crashs Wall Street Flash en 2008, sont dus à des transactions automatisées, avec des machines qui vendent rapidement des actions sans intervention humaine, sur la base de ce qui se passe sur le marché. C'est ce qu'on appelle le trading haute fréquence, dont les divers paramétrages sont de plus en plus maîtrisés par le Machine Learning et l'Intelligence Artificielle. Les experts des données financières utilisent de grandes données pour prédire quels stocks réussiront et quand de futurs crashs sont susceptibles de se produire. Les banques considèrent le Big Data comme un moyen d'augmenter leurs revenus.

Les cartes de crédit et les cartes de fidélité font le suivi de tous les achats en magasin. Ces derniers utilisent des caméras ou même suivent les téléphones portables pour voir quelle partie du magasin retient l'attention des clients le plus longtemps. En ligne, les clients doivent créer des comptes avant de faire des achats, ce qui permet aux sites non seulement de suivre ce qu'ils achètent, mais aussi tous les articles qu'ils consultent. Les magasins basent leurs agencements sur l'intérêt et le comportement des consommateurs. Les vendeurs en ligne décident de ce que nous voyons en fonction des informations démographiques et d'autres critères.

Il y a une grande demande pour le genre de perspicacité qui vient de la surveillance de nos intérêts et de notre comportement en ligne. Facebook et Google sont des géants technologiques rentables en raison de leur capacité à vendre des publicités qui sont mieux à même de cibler des groupes de consommateurs spécifiques que d'autres méthodes et plates-formes publicitaires. Ils peuvent le faire grâce à toutes les informations que les utilisateurs leur fournissent lorsqu'ils utilisent leurs services.

Limites et inconvénients des Big Data

Le Big Data sont prometteuses, mais elles comportent aussi des risques. Premièrement, il y a l'érosion de la vie privée. Plus de gens en savent plus sur chacun d'entre nous qu'à n'importe quel moment de l'histoire de l'humanité. Il n'est pas seulement facile de trouver où nous vivons, mais aussi où nous allons, qui nous aimons, comment nous vivons et ce que nous pensons. Cela rend les individus et les sociétés plus ouverts à la manipulation. Nous pouvons être amenés à donner nos mots de passe et nos numéros de carte de crédit ou à voter pour des candidats que nous n'appuierions pas autrement. Plus de données offrent plus de moyens pour les annonceurs et les entreprises de médias de façonner nos désirs et nos valeurs.

Ensuite, il y a le risque de ce que les gens font avec l'information que les grandes données leur permettent de prédire. Les gens qui ont des habitudes alimentaires malsaines doivent-ils payer davantage pour l'assurance-maladie ? Devrions-nous accroître le maintien de l'ordre dans des domaines où nous prévoyons une augmentation de la criminalité ? Est-ce que nous augmentons les prix pour les acheteurs en ligne qui vivent dans des zones aisées ?

Trouver des moyens d'assurer la sécurité de nos données, le respect de notre vie privée et le maintien de nos valeurs sera un défi permanent au fur et à mesure que la tendance aux données volumineuses se poursuit. Pourtant, peu importe ce que nous en pensons, pour le meilleur ou pour le pire, nous vivons tous dans un monde de grandes données. La cybersécurité est l'un des défis liés à l'essor de l'IoT et du Big Data. Les hackers convoitent de plus en plus les systèmes informatiques. En effet, ils hébergent des données critiques sur lesquelles reposent les objets

connectés. Par exemple, l'émergence des « Smart Cities » ou villes connectées pourraient permettre aux cybercriminels de prendre le contrôle d'une ville entière.

Perspectives et horizons

L'évolution en perspective des Big Data se produit dans un contexte de stratégie Data, caractérisée par un certain self-service Data basé sur la collecte des données pertinentes et exploitables. Cette stratégie fondée sur l'IA et le Deep Learning, facilite la gouvernance des données et ouvrent le chemin devant une amélioration maîtrisée et rentable des produits et des objets des Big Data.

Côté infrastructures, on assiste à une montée en puissance du cloud. Celui-ci présente de très nombreux atouts tels que l'élasticité, la facturation à la demande ou encore la richesse des solutions applicatives. Certaines organisations vont s'orienter vers une approche hybride en hébergeant elles-mêmes les données les plus sensibles et en stockant les autres dans le cloud. Cela permet de tirer parti des atouts des nouvelles architectures tout en assurant le plus haut niveau de protection.

Les bases de données NoSQL et NewSQL (qui offrent des formats de stockage et d'accès aux données performants, agiles, et extensibles) s'imposent et vont poursuivre leur développement dans l'entreprise. Ces innovations permettent de répondre à un enjeu de taille : construire une infrastructure supportant toutes les applications de l'IA à venir, de manière souple et réactive.

Enfin, côté algorithmes, la recherche se concentre sur l'amélioration des capacités d'apprentissage (plus rapide et avec moins de données) et sur la mise au point d'algorithmes de Deep Learning non supervisés.

La gestion de stockage des Big Data s'articule autour des technologies suivantes : Apache, Elastic, Google Cloud Platform, Hadoop, MongoDB, Oracle, Saagie, SAP et Teradata. Avec MongoDB par exemple, les données sont modélisées sous forme de documents JSON, le système permet de faire évoluer le schéma de la base de données à la volée et de s'abstraire de l'utilisation d'un

ORM. Le modèle de type document réduit au maximum le nombre de relation dans la base de données, ce qui simplifie sa structure et augmente sa lisibilité.

Un exemple de code en MongoDb

S'ajoute à ces technologies la Dataviz (Data Visualisation) étant une pratique consistant à mettre en image des données brutes, les rendant ainsi plus accessibles et compréhensibles. Elle donne du sens à ces données. Pour cela, elle fait appel à différentes représentations visuelles. « *La data visualisation, c'est l'art de raconter des chiffres de manière créative et ludique, là où les tableaux Excel échouent. C'est en quelque sorte mettre en musique l'information chiffrée* », explique Charles Miglietti, expert en visualisation de données et co-fondateur de Toucan Toco, éditeur de logiciel de dataviz. Elle apporte ainsi différents avantages précieux aux entreprises.

Constat et observation

Le Règlement général sur la protection des données (GDPR) de l'Union Européenne a été qualifié de « vie privée par défaut », donnant aux citoyens un contrôle strict de leurs données. Plus tôt en 2018, la Californie a adopté la California Consumer Privacy Act (CCPA) pour protéger la vie privée en ligne et les données personnelles Identifiables (DPI). Aujourd'hui, la Federal Data Strategy relative à l'accès aux informations vise à faire de la gouvernance éthique l'un de ses principes fondamentaux. Les grandes entreprises se rejoignent ainsi pour défendre la protection de la vie privée dans l'espoir de façonner la législation future aux États-Unis.

La gouvernance de la donnée est ainsi un ensemble de processus assurant que les données sont formellement gérées à travers l'entreprise. Elle garantit la fiabilité des informations utilisées pour les processus métier critiques, la prise de décision et la comptabilité.

L'intelligence artificielle transforme en profondeur notre société et pose de nombreuses questions éthiques. Les algorithmes reproduisent nos comportements et peuvent induire des biais cognitifs valorisant plus les perceptions que les réalités. Complexes, non fondées sur des règles explicites, les processus et les résultats issus de l'intelligence artificielle sont parfois opaques. L'adoption d'une intelligence artificielle éthique, juste et transparente est indispensable pour son déploiement.

Références :

- https://www.hbrfrance.fr/chroniques-experts/2014/05/2273-quelles-applications-concretes-pour-le-big-data/
- https://web-tech.fr/quest-ce-que-le-big-data-et-ses-applications/
- https://www.next-decision.fr/editeurs-big-data/mongo-db
- https://fr.wikipedia.org/wiki/NoSQL
- https://www.datamation.com/big-data/big-data-technologies.html
- https://fr.blog.businessdecision.com/3-raisons-de-miser-sur-une-gouvernance-des-donnees-efficace/
- https://fr.blog.businessdecision.com/data-7-sujets-chauds-pour-2019/
- https://www.orange-business.com/fr/magazine/predictions-data-2019
- https://fr.blog.businessdecision.com/revue-presse-data-digital-aout-2019/
- https://fr.blog.businessdecision.com/data-storytelling-schisme-business-intelligence/
- https://fr.blog.businessdecision.com/la-dataviz-visualisation-de-donnees-en-lumiere/
- https://www.solutions-numeriques.com/comprendre-le-jargon-du-big-data-explique/
- https://www.lebigdata.fr/iot-big-data
- https://www.lebigdata.fr/fast-data-definition
- https://www.redhat.com/fr/topics/big-data
- https://www.synox.io/4-choses-a-savoir-sur-linternet-des-objets/

- https://www.lebigdata.fr/teradata-solutions-analytiques-entreprise-2608
- https://www.digitalcorner-wavestone.com/2019/05/big-data-quelle-ethique-dans-lanalyse-des-donnees/
- https://www.lesechos.fr/idees-debats/cercle/intelligence-artificielle-pourquoi-leurope-a-sa-carte-a-jouer-1124917
- https://dzone.com/articles/why-is-big-data-in-buzz
- https://www.informatiquenews.fr/peut-on-vraiment-lier-lethique-et-le-big-data-jean-pierre-boushira-veritas-technologies-60788