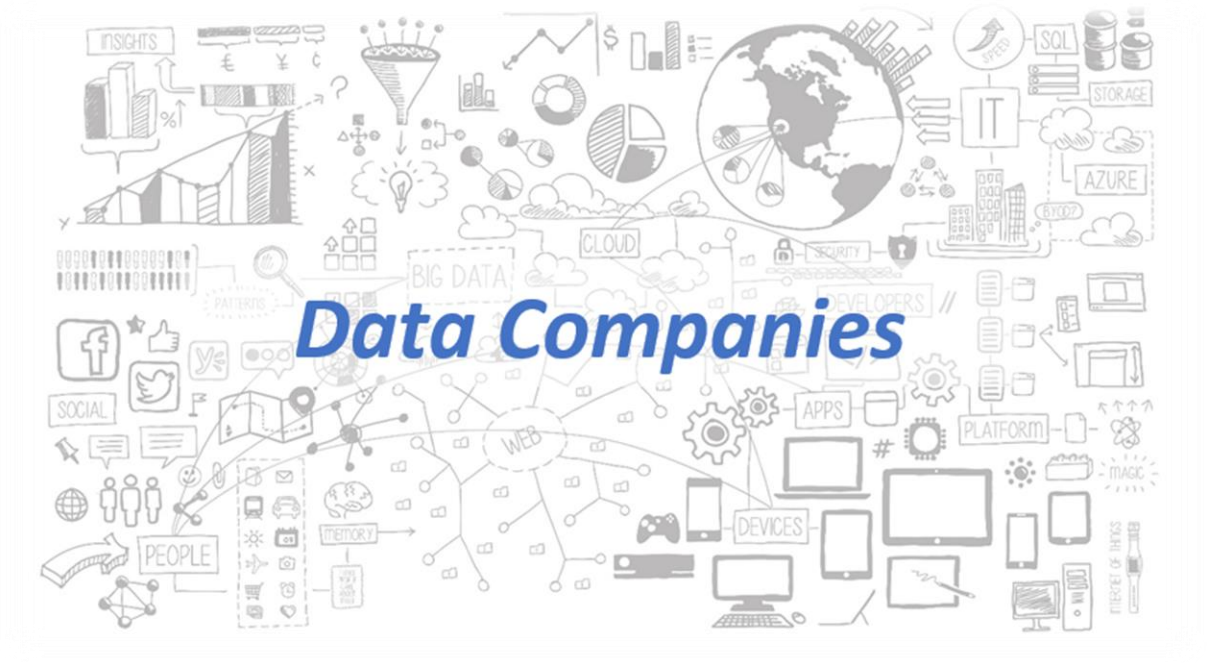
CentraleSupélec

Microsoft



Présenté par
Romain Casteres

Date
Octobre 2018 à Avril 2019

Pour obtenir le diplôme de
RNCP Expert en ingénierie numérique

Table des matières

Genèse du mémoire : C'est en regardant l'intervention du directeur de la technologie (CTO) de la SNCF lors des Microsoft Expériences 2018 dans laquelle il évoque le terme « Data Company » que je me suis intéressé à ce terme. Cette présentation offre un excellent point de départ pour mener une enquête sur ce terme qui semble se répandre dans toutes les sociétés, mais sur lequel on ne sait finalement que peu de choses, notamment sur les conditions de sa mise en œuvre concrète.

Introduction

À l'heure où les données sont omniprésentes, certaines sociétés se disent être des **Data Companies, Data Driven Companies, entreprises pilotées par les données**, etc... Pourtant utilisé à l'origine dans le monde des Startups ou dans les sociétés où la donnée est la compétence principale, aujourd'hui ce terme est cité de plus en plus par les grands groupes.

> *"Data is going to explode across the network", "We are a data company. We are a company that will grow off this data explosion over the next four to five years".* Brian Krzanich (CEO) of Intel (2017).

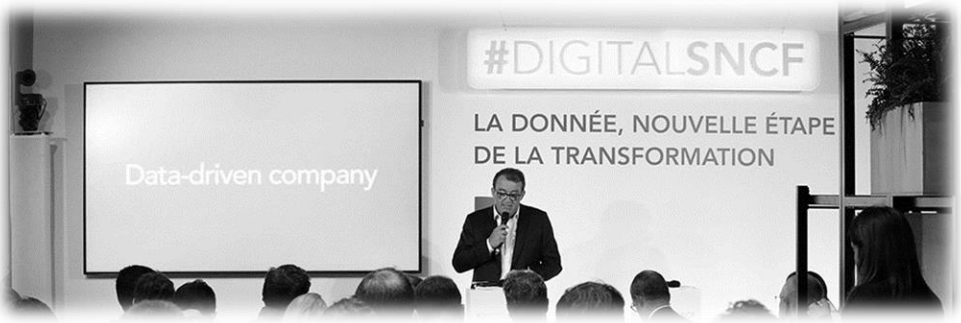


Figure 1 - Benoît Tiers Directeur général e.SNCF, 29 août 2018 : La donnée, nouvelle étape de la transformation de SNCF

Est-ce un effet de mode ou est-ce une nécessité dans le cadre de leur transformation numérique d'intégrer au centre de leur culture la donnée même si cela n'est pas le cœur de leur métier ?

Rappelons que **la transformation numérique est une nécessité pour toutes les entreprises**, leur survie en dépend puisque le numérique améliore l'expérience client et optimise la productivité. **Si la donnée est bel et bien au cœur de transformation numérique, alors elle représente un facteur indéniable dans la conduite de celle-ci.**

Nous tâcherons dans ce mémoire de présenter l'évolution dans l'usage des données numériques (Chapitre 1).

Nous verrons quels sont les impacts des données dans l'entreprise et comment ces entreprises peuvent tirer parti des révolutions technologiques et de leur patrimoine numérique pour innover dans un contexte tiraillé par les contraintes règlementaires (Chapitre 2).

Nous tâcherons ensuite de définir le terme « Data Company » et présenterons quelques mises en œuvre. Enfin, nous verrons à travers un exemple concret comment une entreprise a su se transformer autour d'une architecture data centrique afin de perdurer et gagner des parts de marché (Chapitre 3).

À travers ce mémoire, nous tâcherons de répondre à la question de recherche suivante : **La donnée est-elle au cœur de la transformation numérique des sociétés ?**

1. Omniprésence des données

1.1 – Un peu d'histoire

Historiquement les données étaient rares, le stockage des informations s'effectuait uniquement dans nos cerveaux et sur papier. Les informations étaient difficiles à rassembler, à traiter et à analyser ; sans compter le coût associé au traitement de ces données qui était alors manuel.

Avec l'arrivée du stockage électronique et de la microélectronique, il fut plus facile de stocker et traiter ces informations. La saisie de ces informations fut dans un premier temps réalisée par les êtres humains : stockage des informations financières dans des ERP[1], saisie des bons de commande, des factures, des paiements, des transactions électroniques, …

La quantité de données était alors limitée, mais l'adoption constante de systèmes d'information dans les organisations a généré de plus en plus de données. La plupart de ces données étaient conservées en silo au sein des organisations et leur analyse passait par des rapports que les employés exportaient et imprimaient.

Dans les années 1990, la construction d'entrepôts de données a permis la consolidation de données issues de différents systèmes verticaux et ainsi de fournir aux bonnes personnes les bonnes données pour qu'elles puissent prendre les bonnes décisions.

C'est avec l'arrivée d'internet, des smartphones, des pages Web et des applications mobiles que l'accélération de la création de contenus numériques a explosé. Chaque fois qu'un utilisateur accède à un site Web, les logs de son activité sont sauvegardés. On parle alors du **Big Data** pouvant être défini par les quatre V : **Volume**, **Variété**, **Vitesse** et **Véracité**.

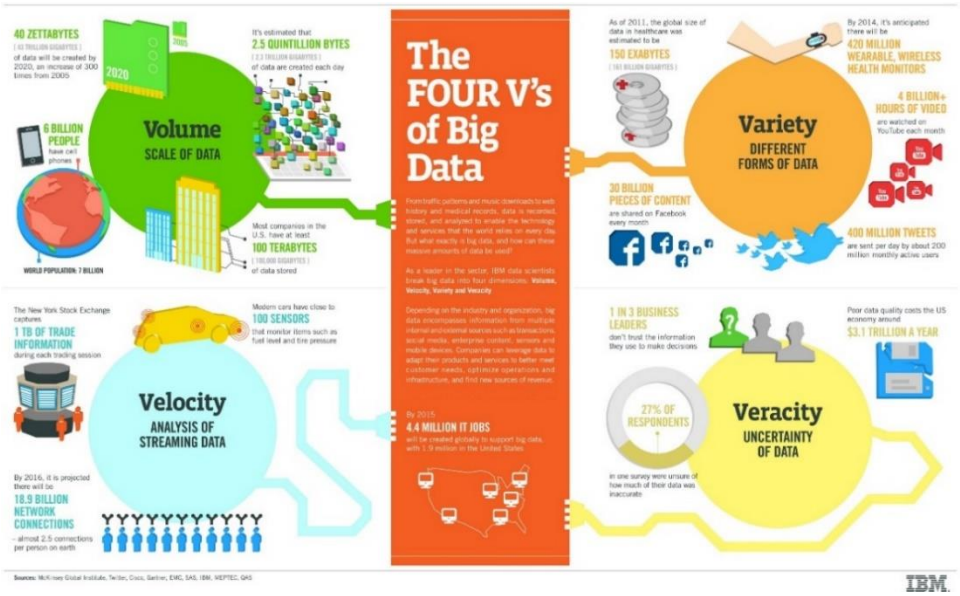
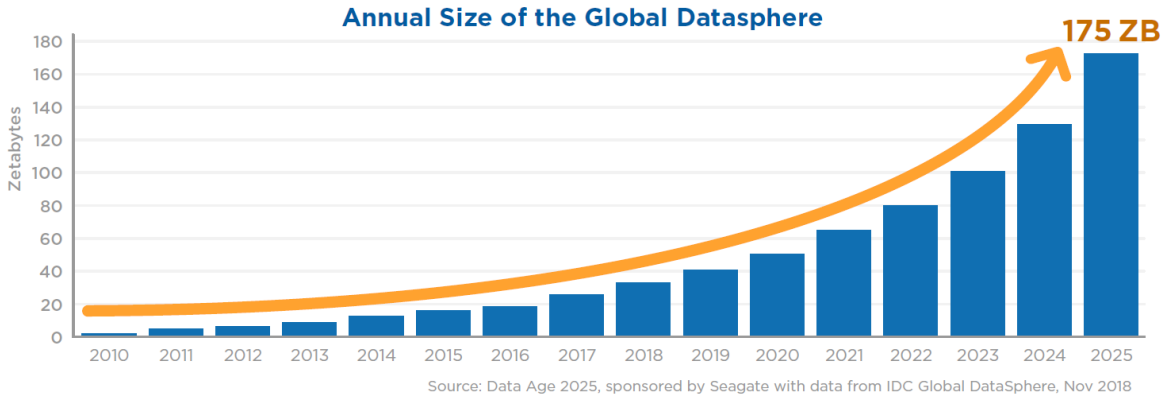


Figure 2 - The Four V's of Big Data, IBM

[1] Enterprise Resource Planning (ERP) : Progiciel permettant de gérer l'ensemble des processus opérationnels d'une entreprise en intégrant plusieurs fonctions de gestion.

Aujourd'hui, le volume de données grossit exponentiellement, des algorithmes de plus en plus sophistiqués sont développés et la puissance de calcul et de stockage est régulièrement améliorée. La convergence de ces tendances alimente les progrès technologiques et les disruptions business.

IDC (International Data Corporation) prédit que le volume de données passera de 33 zettaoctets (ZB) en 2018 à 175 ZB d'ici 2025 :



Les activités visant à générer de la valeur étaient manuelles puis elles furent automatisées et désormais elles sont numérisées.

L'un des principaux résultats de ce processus de numérisation est la production de grandes quantités de données, et ce dans tous les domaines.

Nous assistons à l'explosion des données générées par l'informatique omniprésente, le défi des entreprises est alors de trouver des moyens simples et rapides pour les transformer en sources d'informations.

1.2 – Donnée, Information et Connaissance

Les concepts de « Donnée », « Information » et « Connaissance » revêtent une importance capitale dans de nombreux domaines et notamment dans le monde de l'informatique et des systèmes d'information. Aussi, attardons-nous sur ces termes et précisons ce qu'est une « donnée » et ce qui la différencie de « l'information » ou encore de la « connaissance ».

Étymologie :

- Donnée : Traduction dans le domaine de l'informatique du mot anglais « data », le mot provient du latin « datum », qui signifiait à l'origine (XIII^{e} siècle) « don », « aumône ». Il s'est peu à peu spécialisé en mathématiques, statistiques, psychologie et informatique pour désigner ce qui constitue la base d'un savoir en construction.
- Information : Le mot remonte au XIVe siècle dans le sens de « renseignement que l'on obtient de quelqu'un » et, par extension, désigne au pluriel l'ensemble des connaissances sur un sujet donné. Il vient du verbe « informer », qui signifiait à l'origine « donner une forme ». Il s'emploie généralement en ce qui concerne l'éducation, l'instruction ou toute autre forme de communication du savoir.
- Connaissance : Le mot provient du verbe latin « Cognoscere », qui signifiait « apprendre, se constituer un savoir ».

Par définition, une donnée est un élément brut, qui n'a pas encore été interprété, mis en contexte. Les **données** sont des faits ou des chiffres ; individuellement, elles sont rarement utiles seules, car sans contexte.

Exemple : 10 °.

Lorsque les données sont traitées, interprétées, organisées, structurées ou présentées de manière à être significatives ou utiles, elles sont appelées **informations**. Contextualisées, elles prennent de la valeur.

Exemple : 10°

➔ La température est de 10°C à Paris aujourd'hui.

La présence d'informations ne suffit pas pour prendre des décisions. Ces dernières doivent être interprétées par le cerveau humain pour être transformées en **connaissances** et mener à une action.

Exemple : 10°

➔ La température est de 10°C à Paris aujourd'hui.

➔ Je suis à Paris aujourd'hui donc je m'habille chaudement.

Figure 3 - BALMISSE, Gilles : La recherche d'information en entreprise, Page 68.

Une donnée est *« un élément défini et isolable qui va pouvoir être manipulé, traité et analysé en fonction d'un objectif ou d'un cadre d'analyse ».* [1]

Nous parlerons dans la suite de ce document d'activation de la donnée comme étant l'étape clé permettant la prise de connaissance à partir des données brutes.

L'activation des données consiste à mettre en action les données stockées dans une plateforme de gestion de données, notamment en les traitant et en améliorant leur qualité[2], en d'autres termes, en les rendant intelligibles. Les données sont comme du carburant dans une voiture et l'activation des données est l'allumage, avant le démarrage du moteur la voiture a peu d'intérêt tout comme les données non activées.

« Information usually implies data that is organized and meaningful to the person receiving it. Data is therefore raw material that is transformed into information by data processing. Information can be defined in terms of its surprise value. It tells the recipient something he did not know ». [2]

Une donnée est constituée à partir de règles ou catégories que le producteur de données définit (l'âge, le revenu, les goûts...) ou encore grâce à une unité de mesure communément admise (le mètre, l'euro, le degré...).

Quant à l'information, elle naît de la relation entre une donnée et une personne. Une information n'existe que si une personne interprète une donnée et lui confère du sens. En d'autres termes, les données sont des éléments basiques qui, en fonction du contexte, seront compris et traduits en connaissances par des hommes ou des machines.

Considérer les données comme un capital que l'on peut faire fructifier par des traitements analytiques (statistiques et informatiques) permet de mieux connaître le fonctionnement de l'entreprise et d'améliorer, d'automatiser la prise de décision et les processus. Comprendre les données et les activer dans le but de générer de l'information constitue la base de la « culture »[3] des données.

[2] La qualité des données, en informatique se réfère à la conformité des données aux usages prévus, dans les modes opératoires, les processus, les prises de décision, et la planification (J.M. Juran).

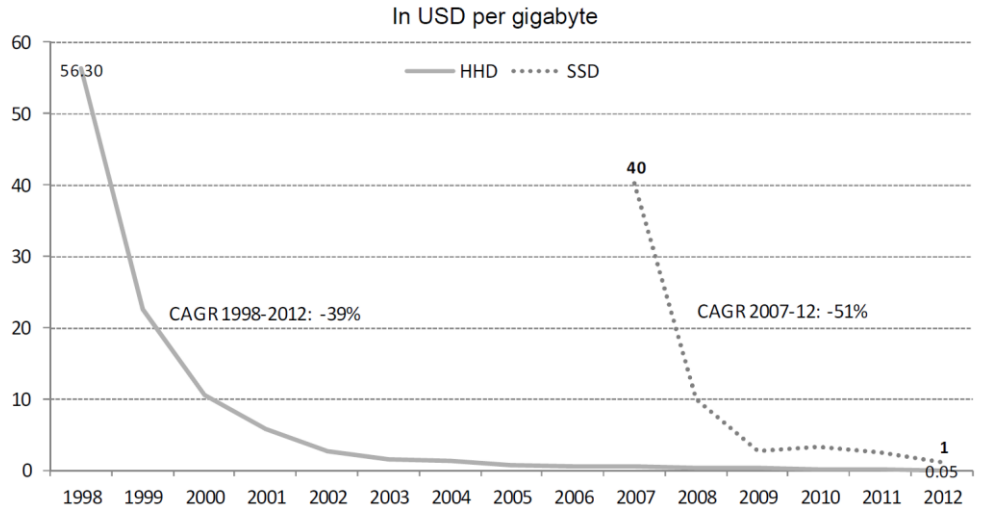
[3] « *La culture des données croise culture médiatique et culture informatique et mathématique. Elle repose comme toute culture sur le partage de connaissances et de pratiques. Il s'agit d'une part de connaître le mode de fonctionnement des médias informatisés qui, entre autres, collectent, communiquent, traitent des données. Regarder et configurer les paramètres de nos applications, comprendre la manière dont est produite une datavisualisation (représentation graphique de données). Et d'autre part, cela requiert d'être en mesure de mobiliser quelques bases informatiques et statistiques* ». [3]

2. Évolution et impacts

2.1 – Le coût des données

Dans le passé, le coût de stockage des données pouvait décourager la conservation de données non utiles. Aujourd'hui les coûts de stockage ayant diminué, les données peuvent généralement être conservées pendant de longues périodes, voire indéfiniment. Ceci est illustré, par exemple, par le coût moyen par gigaoctet de disques durs grand public, qui sont passés de 56 USD en 1998 à 0,05 USD en 2012, soit une baisse moyenne de près de 40% par an. Avec les technologies de stockage de nouvelles générations telles que les disques SSD (Solid-State Drive), la baisse des coûts par gigaoctet est encore plus rapide.

Coût moyen de stockage des données pour les consommateurs entre 1998 et 2012 :



Note: Data for 1998-2011 are based on average prices of consumer-oriented drives (171 HDDs and 101 SSDs) from M. Komorowski (www.mkomo.com/cost-per-gigabyte), AnandTech (www.anandtech.com/tag/storage) and Tom's Hardware (www.tomshardware.com/). The price estimate for SSD in 2012 is based on DeCarlo (2011) referring to Gartner.

Source: OECD based on Pingdom (2011).

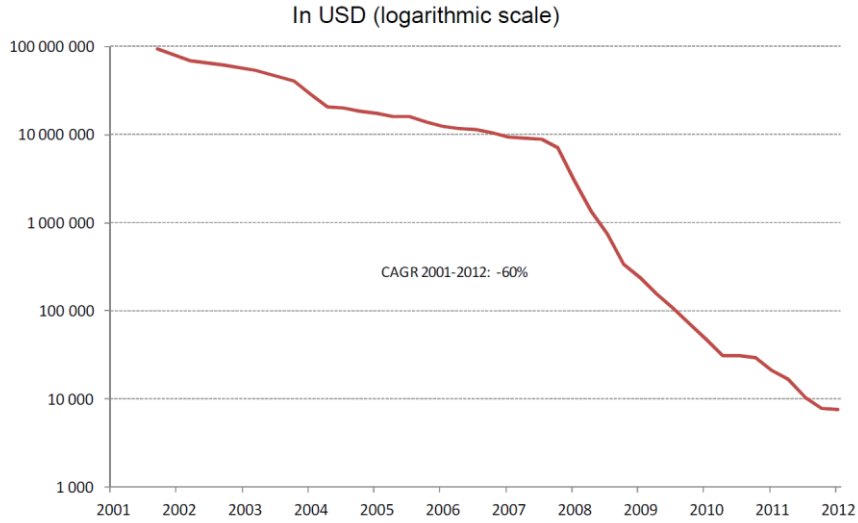
Le CAGR (Compound Annual Growth Rate) est le taux de rendement requis pour qu'un investissement grandisse en supposant que les bénéfices ont été réinvestis.

Parallèlement à la baisse du coût de stockage, la loi de Moore[4], selon laquelle la puissance de traitement double environ tous les 18 mois, a été largement confirmée. Cela est particulièrement visible dans les outils de traitement de données qui sont devenus de plus en plus puissants, sophistiqués, omniprésents et peu coûteux, rendant les données facilement analysables.

En génétique, par exemple, les machines de séquençage des gènes à ADN peuvent maintenant lire environ 26 milliards de caractères du code génétique humain en moins d'une minute et le coût de séquençage par génome a chuté de 60% par an en moyenne, passant de 100 millions USD en 2001 à moins de 10 000 USD en 2012.

[4] La loi de Moore est une loi empirique qui a trait à l'évolution de la puissance de calcul des ordinateurs et de la complexité du matériel informatique.

Coût de séquençage par génome entre 2001 et 2011 :



Source: OECD based on United States National Human Genome Research Institute (www.genome.gov/sequencingcosts/).

Par ailleurs, les infrastructures de communication ont également grandement progressé ce qui a facilité, par la même occasion, l'usage du Cloud Computing lequel a lui aussi joué un rôle important dans l'augmentation de la capacité de stockage et de traitement des données. Nous avons assisté à l'innovation et à la prolifération de dispositifs de production de données qui tirent parti de ces avancées technologiques.

Malgré un coût de transfert de données plus rapide et moins cher ainsi qu'un coût de stockage et de traitement diminué, le cycle de vie des données ne reste cependant pas gratuit. Tout au long de leur cycle de vie, de leur création, en passant par leur transformation jusqu'à leur analyse, les données génèrent des coûts et ce n'est que lorsqu'on commence à les utiliser, les analyser que celles-ci prennent de la valeur.

Coût et valeur de la donnée durant son cycle de vie :

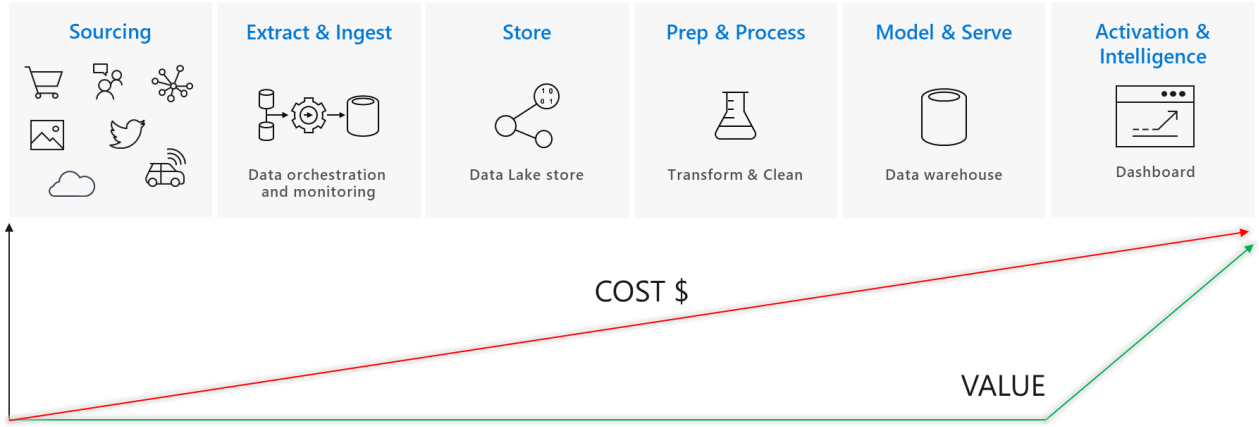
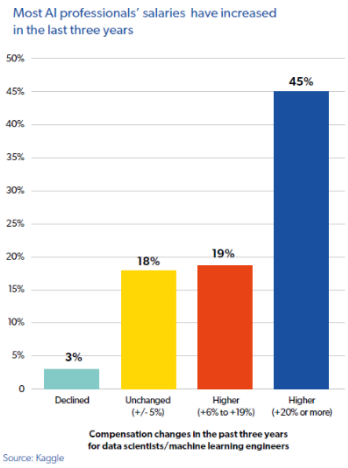
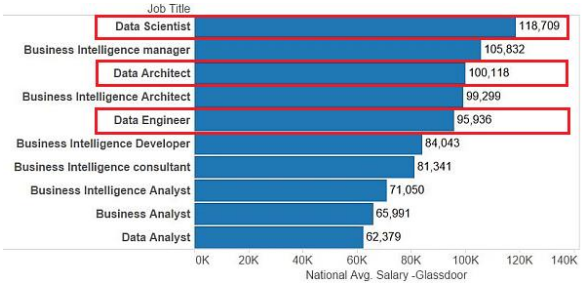


Figure 4 - Coût et valeur de la donnée

Ces différentes étapes nécessitent des infrastructures et des logiciels que des personnes formées doivent savoir utiliser. Ces compétences sont aujourd'hui très recherchées sur le marché du travail et elles ont aussi un coût.

Salaires moyens en dollars aux États-Unis des différents métiers autour de la Data en 2016 :





Sans surprise, dans une étude LinkedIn de 2018 on retrouve ces métiers dans la liste des plus difficiles à recruter et dont les entreprises ont le plus besoin [4]. De plus, une étude du site Kaggle montre que les salaires de certains postes comme celui des Data scientist explosent [5].

Un autre poste de coût à ne pas oublier est celui permettant d'assurer la qualité de la donnée. Une étude de marché réalisée par Gartner en 2018 révèle qu'une qualité médiocre de données entraîne des pertes moyennes de 15 millions de dollars par an [6]. **Si l'analyse des données peut être réalisées sur des données inexactes, incomplètes, ambigües et de qualité médiocre, on peut s'interroger sur le sens à donner aux résultats de ces analyses et remettre en cause, à juste titre, la qualité des connaissances ainsi élaborées**.

Outre les coûts d'activation de la donnée, les salaires des employés, et les traitements pour améliorer la qualité des données, il y a en amont des investissements à réaliser et, pendant toute la durée du cycle, des coûts pour permettre le maintien en condition opérationnelle, la maintenance et la sauvegarde des données. Sans oublier le coût lié à la sécurité et au suivi des règlementations : A titre d'exemple, le RGPD[5] impose aux sociétés travaillant avec l'Europe d'avoir des processus en place permettant la suppression des données de ses utilisateurs en cas de demande. Si le coût d'une mise en conformité efficace (le montant est variable en fonction de la taille de l'organisme, du secteur concerné et de la nature des traitements des données opérés) peut certes être élevé, on peut légitimement rétorquer que ce coût est nettement inférieur à celui des sanctions auxquelles ces sociétés s'exposent en cas de non-conformité. **Les données sont des actifs précieux et une non-conformité peut engendrer des graves conséquences aussi multiples que variées** : sanctions administratives, action en responsabilité, atteinte à la réputation et à l'image de marque, perte de revenus, etc.

La baisse des coûts de la chaîne de valeur des données a été un facteur important de la production et de l'utilisation croissante de données ainsi que de la migration accélérée des activités socio-économiques vers Internet grâce à l'adoption généralisée des services électroniques dans un environnement de plus en plus participatif. Cependant, même si les coûts de l'activation des données ont baissé, ces activités requièrent des connaissances, des outils, des infrastructures ainsi que le respect des règlementations, … générant des coûts ; le retour sur investissement ne sera possible que quand les données auront été activées.

[5] RGPD : Le Règlement Général sur la Protection des Données, est un règlement de l'Union européenne qui constitue le texte de référence en matière de protection des données à caractère personnel. Il renforce et unifie la protection des données pour les individus au sein de l'Union européenne.

2.2 – La valeur des données

La numérisation et la Datafication[6] touchent tous les secteurs de l'économie et créent un vecteur continu de changements dans l'économie. La valorisation de nombreuses entreprises telles que Google, Amazon et Facebook repose aujourd'hui sur deux facteurs : Leur potentiel à générer des profits à partir de leur base d'utilisateurs et la quantité de données qu'elles gèrent.

Par exemple, en 2015, Facebook valait environ 250 milliards de dollars tandis qu'Air Canada, une société qui détient des actifs corporels tels que des avions et détient des licences lui permettant d'utiliser des installations aéroportuaires et d'opérer dans le monde entier, ne valait que 34 milliards de dollars.

Si les données peuvent aider à créer de nouveaux produits ou gammes de services, elles peuvent en revanche constituer un obstacle à l'entrée sur le marché pour les nouvelles entreprises confrontées à une concurrence bien établie. Par exemple, si quelqu'un devait créer demain un nouveau service de réseau social professionnel, il serait impossible pour lui de rivaliser avec LinkedIn dans la mesure où LinkedIn a tellement de données sur ses utilisateurs professionnels que sans ces données, il aura beaucoup de mal à entrer sur le marché.

On entend souvent *« La Data est le pétrole du XXIe siècle »,* certaines entreprises ont intégré cette certitude depuis longtemps, d'autres commencent seulement à comprendre tout l'avantage stratégique qui en découle. Avec la multiplication des appareils et objets connectés, des plateformes sociales et surtout des solutions de suivi des internautes, le client se retrouve de plus en plus "créateur de données", données qui vont ensuite pouvoir être qualifiées pour bâtir une stratégie et ainsi générer plus de valeurs.

Un autre exemple illustrant le transfert de l'actif corporel aux actifs incorporels, telles que les données, dans la valorisation et la stratégie d'entreprise : En 2015, IBM a acquis Weather Company pour 2 milliards de dollars. Celle-ci fournit le contenu météo de plus de 195 000 stations. Alors que de nombreux analystes étaient surpris par l'acquisition d'une société de météorologie par une société de technologie appartenant à un secteur totalement différent, un an plus tard, IBM utilisait les données de la société de météorologie pour répondre aux préoccupations de leurs commerçants face aux intempéries avec un modèle de prévision très précis. Aujourd'hui leurs solutions fournissent aux télédiffuseurs, pilotes, commerçants en énergie, agents d'assurance, employés de l'État, responsables de la vente au détail, et plus encore, des informations sur l'impact des conditions météorologiques sur leurs activités, les aidant à prendre des décisions plus éclairées pour améliorer la sécurité, réduire les coûts et générer des revenus.

[6] Le terme Datafication apparu pour la première fois dans le livre "Big Data: A Revolution That Will Transform How We Live, Work and Think" (Kenneth Cukier, 2013) fait référence à l'hyperinflation numérique, qui se traduit par des masses gigantesques et difficilement gouvernables de données. Le processus de Datafication est parallèle au processus de transformation des actifs en actifs numériques et intelligents.

L'infonomie[7] est une théorie axée sur la quantification de la valeur de l'information et la gestion de l'information comme un atout. La portée de l'infonomie consiste à traiter les données avec une monétisation identique ou similaire à celle des autres actifs traditionnels tels que les actifs financiers, physiques ou immatériels. Selon la théorie infonomique, une information répond à tous les critères des actifs traditionnels d'une entreprise. Même si elles ne sont pas encore acceptées par les pratiques comptables les plus courantes, de plus en plus d'organisations se comportent comme si elles devaient optimiser la valeur commerciale générée par l'information. De la même manière, l'économie de données (Data Economy) implique de tirer parti des données et de leur analyse comme atout pour augmenter les revenus de l'entreprise, accroître son efficacité opérationnelle ainsi que générer de nouveaux modèles commerciaux [7]. L'économie de données a également été reconnue par la Commission européenne. Dans un article intitulé *« Towards a thriving data-driven economy »* publié sur le site Web de la Commission Européenne, les auteurs déclarent que l'économie fondée sur les données encouragera la recherche et l'innovation ce qui créera de nouvelles opportunités commerciales. [8]

La monétisation des données consiste à utiliser des données pour obtenir des avantages économiques quantifiables directs (vente via un courtier en données ou indépendamment) ou indirects (améliorations mesurables des performances commerciales, innovations, réduction des risques et des coûts). [9]

Nous l'avons déjà évoqué, la valeur de la donnée n'est véritable que si elle est activée, mais d'autres aspects sont à prendre en compte quant à sa valeur :

- Son volume : Plus le volume de données est grand, plus les informations pouvant en être extraites seront potentiellement importantes, par exemple le fait de pouvoir analyser un plus grand historique de données afin de déceler des patterns, des schémas répétitifs.
- Sa variété : La capacité d'acquérir et d'analyser des données variées (structurées ou non structurées) est extrêmement précieuse. Plus les données clients d'une entreprise seront diversifiées, plus elle sera à même de développer une vision globale et ainsi mieux cibler les attentes de ses clients.
- Sa qualité, sa véracité : Une mauvaise complétude, conformité, cohérence, précision, unité ou encore intégrité dans les données impacte lourdement sa valeur. Par exemple une donnée erronée sur l'adresse d'un utilisateur impactera les coûts d'envois postaux.
- Sa fraicheur : Les données ne se déprécient pas en raison de leur consommation, mais elles peuvent devenir non pertinentes en raison de leur ancienneté. En expliquant la théorie de l'innovation de rupture « Disruptive innovation », Clayton Christensen disait que le problème des paradigmes de type Data-Driven est que les données viennent du passé et qu'on risque de prendre des décisions quand il sera déjà trop tard [10]. Les données concernant le futur n'existent pas et personne ne peut nous les fournir. Pour optimiser, nous avons donc besoin d'exploiter les données du présent tout de suite et rapidement avant qu'elles ne soient « périmées ».
- Sa rareté : Les biens les plus rares ont le plus de valeur. Ce principe a une logique économique indéniable et reste vrai concernant les données. Leur rareté participe à déterminer leur prix.

[7] L'infonomie (Infonomics en Anglais), s'intéresse à la manière de produire, de diffuser et d'utiliser des contenus immatériels : données, informations, connaissances, savoir-faire, etc.

La valeur des données augmente si la capture, l'analyse et le traitement des données variées et de qualités sont rapidement réalisés :

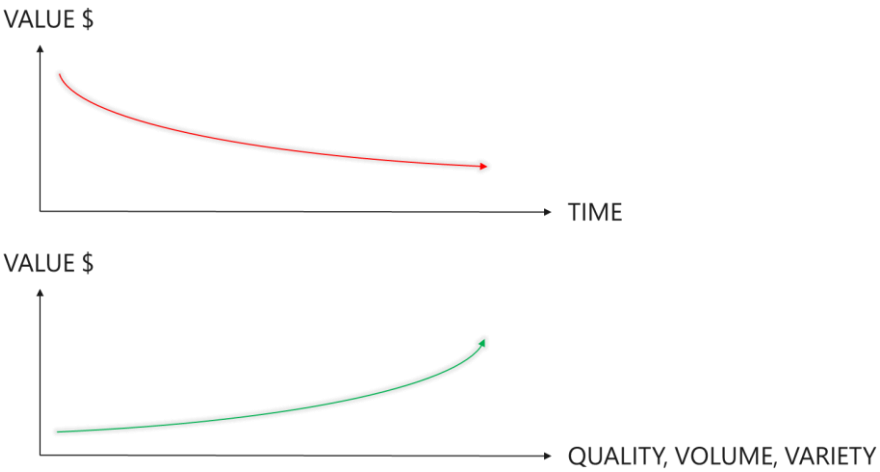


Figure 5 - Valeur de la donnée [8]

Les données historiques porteuses de moins de valeurs à l'instant T sont toutefois porteuses d'informations. Au regard du coût de stockage, il est aujourd'hui plus pertinent de les sauvegarder et de trouver une façon de les activer par la suite plutôt que de les supprimer.

D'autres traitements sophistiqués peuvent ajouter de la valeur aux données. C'est le cas de l'analyse prescriptive et de l'analyse prédictive qui tirent parti de l'apprentissage automatique :

- **L'analyse prédictive** permet de faire des projections futures basées sur des données historiques et actuelles. Elle permet d'utiliser les données brutes connues et de les traiter afin de pouvoir prédire les informations non connues.
- **L'analyse prescriptive** permet non seulement de donner un sens aux données brutes, mais également de s'en servir pour déterminer les actions à prendre**.** L'analyse prescriptive aide également à faire évoluer la logique décisionnelle afin de maintenir ou d'améliorer son efficacité au fil du temps.

Il est courant de présenter les étapes de sophistication de la donnée sous la forme suivante :

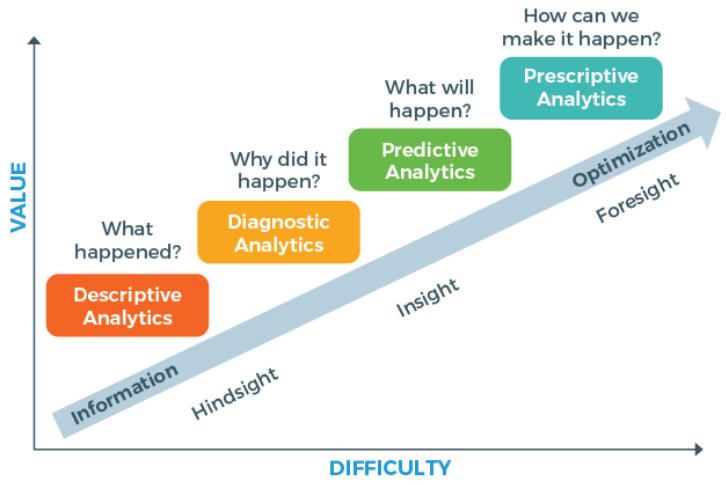


Figure 6 - Analytics Value Escalator [11]

[8] Schéma inspiré de la figure 26 "The value of information and data in time" du livre The Data Advantage [12].

Autre caractéristique intéressante de la donnée : Elle est aujourd'hui facile à distribuer et peut être réutilisée plusieurs fois contrairement à d'autres actifs. Selon G. Shapiro et H. Varian dans « *Information rules* » : « *l'information est coûteuse à produire mais peu coûteuse à reproduire* ». Le prix est fixé en fonction de sa valeur pour le client et non en fonction de son coût marginal. Le PDG de Saint-Gobain a énoncé dans une conférence (Conférence de l'IAE de Paris du 09/05/2016 sur les enjeux de la transformation numérique pour les entreprises) que l'investissement dans une plateforme a un coût élevé au début pour la mettre en place, mais un coût marginal faible par la suite pour exploiter les produits numériques qu'elle génère.

Il n'est donc pas étonnant de voir naitre de plus en plus d'entreprises spécialisées dans la vente de données (**Data Broker**). À titre d'exemple, la société américaine Datalogix (rachetée par Oracle en 2014) traque les Data issues des transactions bancaires transitant par les sites de commerce en ligne ou chez les commerçants et magasins physiques. Ces données amassées sont notamment vendues à de grands groupes comme Facebook ou Google pour les aider à mieux cibler leurs offres publicitaires. En France, la société Dawex est une « place de marché » de données où les entreprises achètent et vendent des données. Dans ces places de marché, la donnée peut avoir plusieurs niveaux de valorisation : sa valeur brute (par sa source ou son contenu propre) et sa valeur « raffinée » en fonction de l'expertise, de son traitement, de sa fiabilité ou encore de sa pertinence.

Comme précisé ci-avant, la donnée a une valeur significative, mais un mouvement prône son ouverture, son partage. Initié aux États-Unis en 2009, ce mouvement a trouvé un fort écho dans la communauté des développeurs notamment parmi les contributeurs de projets Open Source dont le mouvement est philosophiquement proche : L'**Open Data**[9]. L'ouverture des données est à la fois un mouvement, une philosophie d'accès à l'information et une pratique de publication de données librement accessibles et exploitables. Elle s'inscrit dans une tendance qui considère l'information publique comme un bien commun dont la diffusion est d'intérêt public et général. Une donnée ouverte répond à un ensemble de critères techniques, économiques et juridiques : Elle doit être accessible gratuitement et librement en ligne dans un format qui en permet la réutilisation. Des exemples d'Open Data : Données sur le transport, la cartographie, les statistiques, la géographie, la sociologie, l'environnement, etc.

Les données ont une valeur et procurent un avantage concurrentiel certain aux entreprises en capacité de les activer. La valeur des données dépend de plusieurs facteurs : Leur volume, leur variété, leur véracité, leur qualité, leur rareté, leur fraicheur et leur sophistication.

La monétisation des données révolutionne la façon dont les entreprises considèrent les actifs informationnels. Les informations générées à partir des données brutes sont monétisables et servent à obtenir un avantage économique certain. Ainsi la capacité de produire, de comprendre et d'utiliser des données numériques devient une compétence essentielle pour toutes entreprises.

[9] Le terme open data trouve son origine en 1995 dans une publication du comité sur les données géophysiques et environnementales du Conseil national de la recherche aux États-Unis intitulée De l'échange complet et ouvert des données scientifiques.

2.3 – L'innovation pilotée par les données

Compte tenu de la dynamique de l'industrie moderne, l'innovation devient de plus en plus importante pour assurer la croissance, la durabilité et la compétitivité des entreprises.

Le dictionnaire Larousse définit l'innovation comme un *« Ensemble du processus qui se déroule depuis la naissance d'une idée jusqu'à sa matérialisation (lancement d'un produit), en passant par l'étude du marché, le développement du prototype et les premières étapes de la production ».* Ainsi, l'innovation ne désigne pas seulement la nouvelle idée, mais aussi toutes les étapes nécessaires à sa réalisation.

Le manuel d'Oslo[10] définit les quatre formes de l'innovation entrepreneuriale :

- Innovation de produit : Elle signifie l'introduction d'une nouvelle offre, d'un nouveau bien ou service,
- Innovation de procédé (ou de processus) : Elle signifie la production d'un produit selon une nouvelle méthode. C'est par exemple la mise en place d'une nouvelle méthode de production par l'exploitation d'une nouvelle technologie,
- Innovation de commercialisation (ou de marketing) : Elle signifie la production d'un même produit avec la même méthode, mais avec un changement de design ou l'utilisation d'une nouvelle méthode de commercialisation impliquant des changements significatifs de la conception ou du conditionnement, du placement, de la promotion ou de la tarification d'un produit.
- Innovation d'organisation (ou managériale) : Dans ce cas, on fabrique la même chose de la même manière et on la commercialise de la même façon, mais avec une nouvelle organisation plus efficiente. Les méthodes Lean Management et Agile sont des exemples de ce type d'innovation.

Ces quatre formes d'innovation ont un point commun : L'utilisation des technologies émergentes et l'amélioration constante des processus de création de valeur. Les entreprises doivent disposer de souplesse et d'agilité nécessaires pour créer de nouvelles offres, produits ou encore services. Ils adoptent alors des méthodologies comme avec le Lean Startup.

Le Lean Startup est une méthode alternative au développement traditionnel des produits et des business qui a été introduite par Eric Ries en 2011. Contrairement au processus linéaire du développement de produit traditionnel, le Lean Startup est un développement agile avec des cycles courts et répétés. L'objectif principal du Lean Startup est de raccourcir les cycles de développement de produits et de développer des produits conformes aux besoins des clients. La méthode est liée à l'approche Lean Manufacturing en visant à éliminer les activités inutiles et à augmenter les actions à valeur ajoutée dans le processus de développement de la production. L'agilité de la méthode repose sur la réception continue des commentaires des clients tout au long du processus de développement du produit, ce qui permet d'éviter de consacrer des ressources à des fonctionnalités indésirables qui ne répondent pas aux besoins des clients et réduisent les risques du marché.

[10] Le Manuel d'Oslo de l'Organisation de Coopération et de Développement Economiques (OCDE) rassemble les principes directeurs proposés pour le recueil et l'interprétation des données sur l'innovation.

Boucle de feedback « Build-Measure-Learn » :

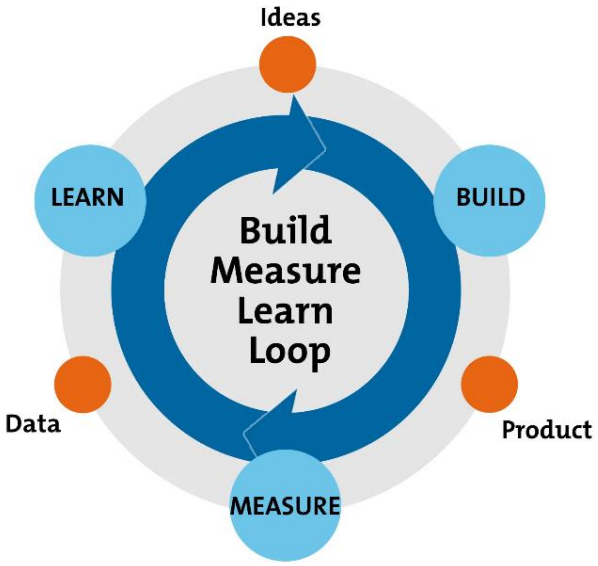


Figure 7 - Diagramme adapté d'Eric Ries (2011) 'The Lean Startup' [13]

Les entreprises dont les compétences en matière de données sont plus avancées extraient de plus en plus de valeurs à partir de ces données et informations et deviennent des entreprises intelligentes.

Exemple avec le robot de cuisine Thermomix fabriqué par la société Vorwerk : Thermomix est sur le marché depuis plus de trente ans et en est actuellement à la cinquième génération de son produit. La proposition de valeur de Thermomix est qu'il permet aux personnes qui ne savent pas ou n'ont pas le temps de cuisiner, de cuisiner. En baisse de vente, la société a incorporé des fonctionnalités dans sa dernière génération qui en ont fait un actif intelligent. En effet, le Thermomix peut désormais se connecter au réseau Wifi du domicile du client et peut ainsi recevoir et envoyer des données. Il peut recevoir des recettes de cuisine et envoyer des informations au réseau sur l'utilisation du produit. Cet ajout intelligent à la plateforme permet au Thermomix de faire beaucoup de choses qu'il ne pouvait pas faire par le passé : tout d'abord, il crée des données sur la manière dont les utilisateurs utilisent le produit, ce qu'ils cuisinent et à quels moments. Ces données peuvent ensuite alimenter le développement futur des produits de la société. En outre, la plateforme envoie des données sur les performances techniques du produit ainsi que sur les éventuels dysfonctionnements. Toute cette information revient au créateur de Thermomix qui en apprend ainsi beaucoup sur les habitudes d'utilisation ou encore les goûts de ses clients. De plus, Thermomix dispose d'une équipe de cuisiniers qui créent constamment de nouvelles recettes pour leurs utilisateurs. En connaissant les habitudes de ses utilisateurs, Thermomix a la possibilité de leur envoyer des recommandations personnalisées. Cet exemple montre clairement le potentiel des actifs intelligents (pertinents pour de nombreux autres actifs tels que les maisons intelligentes, les montres, les voitures ou les portes), ainsi que la valeur des produits de données et la manière dont ils améliorent les actifs numériques et les actifs intelligents.

L'innovation pilotée par les données ou en anglais **Data-Driven-Innovation** (DDI) a été définie et présentée dans le livre « Exploring Data-Driven Innovation as a New Source of Growth » [14].

A la lecture de cet ouvrage, il est clair que les innovations de demain seront en grande partie pilotées par les données :

> *« DDI has the potential to enhance resource efficiency and productivity, economic competitiveness, and social well-being as it begins to transform all sectors in the economy, including low-tech industries and manufacturing. The exploitation of DDI has already created significant value-added for many businesses and individuals, and more can be expected to follow. […] Available evidence also shows that firms using DDI have raised productivity faster than non-users by around 5-10%. ».*

Différents exemples de l'utilisation des données comme source d'innovation et de croissance de la productivité :

- Utilisation des données pour la création de nouveaux produits (biens et services). Cela inclut l'utilisation de données en tant que produit (« Data Products ») ou en tant que composant majeur d'un produit (produits à forte consommation / création de données « data-intensive products »).
- Utilisation des données pour optimiser ou automatiser les processus de production ou de livraison (processus pilotés par les données, « Data Driven Processes »). Cela inclut l'utilisation des données pour améliorer l'efficacité de la distribution des ressources énergétiques (réseau électrique intelligent, « Smart Grids »), de la logistique et du transport (logistique et transport intelligent).
- Utilisation des données pour améliorer le marketing, par exemple en fournissant des publicités ciblées et des recommandations personnalisées ou d'autres types de clustering liés au marketing (marketing fondé sur les données, « Data Driven marketing), ainsi que l'utilisation de données pour la conception de produits expérimentaux (conception de produits axée sur les données, « Data Driven Product Design »).
- Utilisation des données pour de nouvelles approches organisationnelles et de gestion ou pour améliorer de manière significative les pratiques existantes (organisation et prise de décisions basées sur des données, « Data Driven organization and Data Driven Decision Making »).
- Utilisation des données pour améliorer la recherche et le développement (« Data Driven R&D »).

Les différentes formes d'innovation pilotées par les données s'inscrivent naturellement dans la « boucle de feedback » permettant aux sociétés d'améliorer en continu leurs produits, procédés ou encore leurs organisations et ainsi de proposer de nouveaux services à leurs clients et prospérer.

D'autres sociétés utilisent les produits de données comme source de diversification pour créer de nouveaux business ou encore dans le but d'améliorer des processus existants.

Ces sociétés ont toutes un point commun, celui de mesurer, d'apprendre et ainsi de progresser via l'utilisation des données.

2.4 – Les transformations liées à la donnée

Nous avons vu le coût de la donnée, quelle pourrait en être la valeur si celle-ci était activée et comment les sociétés peuvent les utiliser pour innover. Etudions à présent les transformations nécessaires afin d'en tirer parti en tant qu'actif stratégique.

a. Technologie

Afin de générer des informations numériques, les activités de l'entreprise doivent être numérisées. Être une entreprise numérique est la première étape pour disposer d'actifs numériques et avoir la capacité d'extraire de la valeur de données produites par ces actifs numériques. En raison de la dématérialisation d'un nombre croissant de processus et l'apparition de produits et services numériques totalement nouveaux, la quantité de données numériques créées augmente de manière exponentielle.

L'omniprésence et la prolifération exponentielle des données impactent les entreprises qui doivent déployer des plateformes de gestion et d'exploitation de la donnée capables de répondre aux nouveaux besoins des utilisateurs comme l'analyse en temps réel, l'agrégation d'énormes quantités de données ou encore l'analyse prédictive. Regardons quelques solutions technologiques supportant ces transformations.

La gestion des données d'entreprise (**EDM** en anglais : Enterprise Data Management) est la capacité d'une organisation à créer, intégrer, diffuser et gérer efficacement des données pour l'ensemble des applications, processus et entités de l'entreprise. Une Data Management Platform (**DMP**), ou Plateforme de Gestion des Données permet de collecter et centraliser les données, de les retraiter (en les unifiant, les enrichissant, les segmentant, …) dans le but de les activer. Le Master Data Management (**MDM**) désigne un **ensemble d'outils et de méthodologies** permettant l'intégration et la maintenance des données de références (Master Data) de l'entreprise. Ces outils sont étroitement liés à la gouvernance des données, ils jouent alors un rôle prépondérant dans la transformation liée aux données en permettant notamment de cartographier les données de l'entreprise et d'avoir « une seule version de la vérité ».

Dès l'invention des systèmes de gestion de base de données (**SGBD**), les bases de données relationnelles (**OLTP**, OnLine Transaction Processing) ont présenté, entre autres, l'avantage de retrouver rapidement une donnée particulière grâce aux index. Toutefois, ces mêmes bases de données ne sont pas adaptées aux traitements de masse nécessaires aux calculs des données d'un système décisionnel, où l'objet est de ramener **un grand nombre de données et de les agréger** entre elles. Ainsi les systèmes **OLAP**[11] ont fait leur apparition, ils précalculent (ou du moins facilitent largement les agrégations) toutes les valeurs clés (aussi appelées **mesures**) selon des axes d'analyse (ou **dimensions**).

Datawarehouse : Le système d'information de l'entreprise et notamment son système d'aide à la décision, de Business Intelligence (**BI**) a un rôle primordial dans l'entreprise.

[11] Les systèmes OLAP ont été conceptualisés par Edgar Frank Codd en 1996.

Voici un schéma classique d'un projet décisionnel :

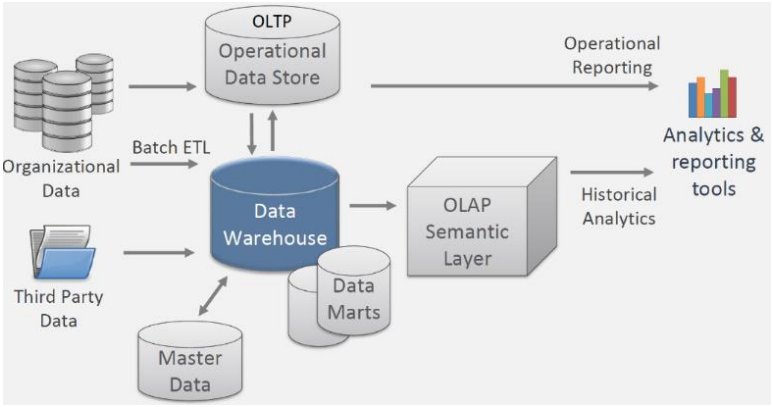


Figure 8 - { Architecture V1 } - Système d'information décisionnelle traditionnelle

Le Datawarehouse est né du besoin de centraliser les données des différentes bases opérationnelles de l'entreprise dans le but d'effectuer des analyses sur plusieurs axes (Dimensions). Le Datawarehouse doit aujourd'hui évoluer afin de pouvoir répondre aux nouveaux usages liés à la transformation numérique et au besoin de réactivité.
Citons ici quelques-unes de ces limitations :

- Le coût de stockage.
- La latence dans l'insertion de nouvelles données dans des schémas bien définis, structurés.
- Les analystes n'ont pas accès aux données brutes d'origine et sont limités à l'utilisation du sous-ensemble extrait de l'entrepôt de données.
- Seul un traitement sur des données structurées est possible dans un entrepôt de données. Aucune application d'apprentissage ou d'analyse utilisant des informations non structurées n'est réalisable.
- Les mises à jour des Datawarehouses sont très lentes, car elles sont en général effectuées par lots (Batchs) journaliers, hebdomadaires ou même mensuels.

Ces problèmes créent de sérieuses limitations pour rendre les données et le traitement accessibles au plus grand nombre.

Data Lake : James Dixon, fondateur et directeur de la technologie de Pentaho, a été le premier à utiliser le terme « Data Lake » en 2010. Selon Dixon, un lac de données est un référentiel de stockage contenant une grande quantité de données brutes dans leur format natif jusqu'à ce qu'elles soient nécessaires.

Les Data Lakes traitent les lacunes des entrepôts de données de deux manières :

- Premièrement, dans les lacs de données, les données peuvent être stockées dans des formats structurés, semi-structurés ou non structurés.
- Deuxièmement, le schéma de données est décidé lors de la lecture des données plutôt que lors de leur chargement ou de leur écriture. Il est ainsi toujours possible de modifier le schéma en présence d'informations supplémentaires ou lorsqu'il est nécessaire de récupérer d'autres informations à partir des données brutes. Il y a alors une amélioration de l'agilité organisationnelle. Cela signifie également que les données sont rapidement disponibles, car elles n'ont pas besoin d'être préparées avant d'être consommées par les moteurs de traitement.

Enfin, en raison de la rentabilité des Data Lakes, il n'est jamais nécessaire de jeter ou d'archiver les données brutes.

Voici un comparatif Datawarehouse et Data Lake :

Data Warehouse	vs.	Data Lake
structured, processed	**Data**	structured / semi-structured / unstructured, raw
schema-on-write	**Processing**	schema-on-read
expensive for large data volumes	**Storage**	designed for low-cost storage
less agile, fixed configuration	**Agility**	highly agile, configure and reconfigure as needed
mature	**Security**	maturing
business professionals	**Users**	data scientists et al.
SQL	**Types of Queries**	Multiple types: programmatic access, machine learning, SQL, graph analysis, deep learning, etc.
mature	**Tool Integration**	maturing

Figure 9 - Datawarehouse vs Data Lake

En résumé, **les Data Lakes offrent beaucoup plus de flexibilité et d'agilité, ce qui est essentiel pour une entreprise basée sur les données**. Cela pour un coût généralement plus bas que le Datawarehouse.

De ce fait de nouveaux termes spécifiant les « Modern Datawarehouse » sont apparus, par exemple, Mark Beyer du Gartner introduit en 2011 le terme **Logical Data Warehouse** (LDW) et le définit ainsi « *Une nouvelle architecture de gestion des données pour l'analyse combinant les atouts des entrepôts de données traditionnels avec des stratégies alternatives de gestion et d'accès aux données* » [15].

Big Data Framework, Hadoop, Spark : Le Big Data n'est pas une nouvelle discipline basée sur une seule technologie, mais une combinaison d'anciennes et de nouvelles technologies permettant de gérer une quantité énorme de données.

Développé par Apache Software Foundation, Apache Hadoop est un Framework Java open source destiné à faciliter la création d'applications distribuées et Scalables. Le Framework est composé d'un système de fichiers distribués Hadoop Distributed File System (**HDFS**) sur lequel on exécute des traitements parallèles appliquant le principe de **Map & Reduce**. Hadoop propose une bibliothèque de logiciels permettant d'extraire, de transformer, d'analyser, … les données. En voici quelques-uns :

- Pig : Framework permettant d'exprimer dans un langage de plus haut niveau (Pig Latin) ce qui se traduira automatiquement par des tâches Map/Reduce.
- Hive : Même principe que Pig, le langage (HiveQL) est quant à lui très proche du SQL.
- Mahout : Bibliothèques pour le Machine Learning.
- Pegasus : Permet d'analyser des pétaoctets de données contenues dans des graphes.
- Sqoop : Permet de transférer des données entre Hadoop et des SGBDR.
- Flume : Un service fiable, efficace et disponible pour la collecte, le regroupement et le déplacement de grandes quantités de logs.

De même, Apache Spark est un Framework open source de calcul distribué. En prenant en charge le In-Memory (l'informatique en mémoire), Spark est en mesure d'interroger les données beaucoup plus rapidement que les moteurs basés sur disque tels que Hadoop. Spark n'a pas été prévu pour remplacer Hadoop, mais pour mettre à disposition une solution complète et unifiée permettant de prendre en charge différents cas d'utilisation et besoins dans le cadre des traitements Big Data.

On retrouve dans Spark les outils suivants :

- Spark Streaming : Utilisé pour le traitement en temps réel des données.
- Spark SQL : Utilisé entre autres pour extraire, transformer et charger des données sous différents formats et les exposer pour des requêtes ad-hoc.
- Spark MLlib : C'est une librairie de Machine Learning qui contient des algorithmes et utilitaires d'apprentissage comme la classification, la régression, le clustering, ...
- Spark GraphX : Utilisé pour les traitements de données stockées dans des graphes

Le paysage autour des technologies Big Data ne cesse de s'agrandir :

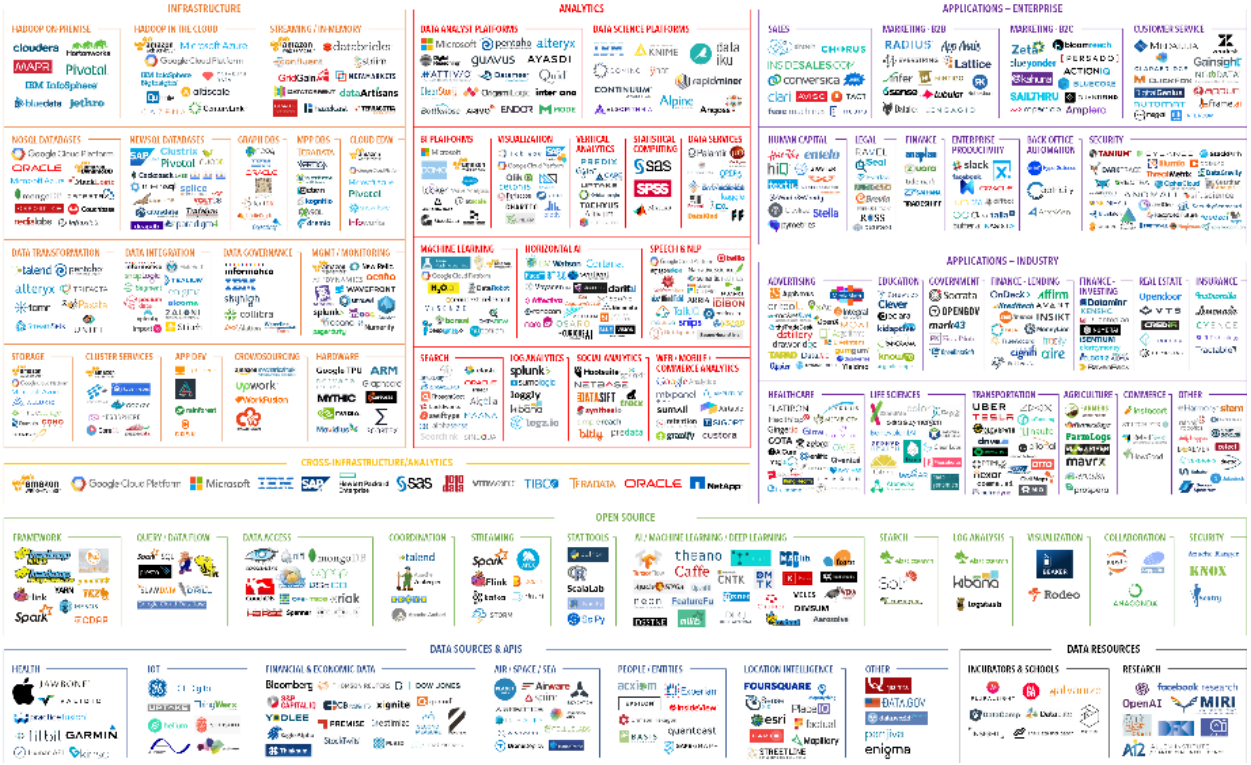


Figure 10 - Big Data Landscape 2017

Ces Frameworks d'analyse de données volumineuses apportent également de nouvelles générations de bases de données et de nouveaux modèles de programmation conçus pour gérer les données volumineuses de manière plus efficace.

No-SQL : Dans les systèmes de bases de données relationnelles, la structure des données est définie au moment du design sous la forme de tables, ce qui signifie qu'elle est conçue avant le chargement des informations dans le système. La structure de données comprend le modèle relationnel, la structure de table, la largeur de colonne et les types de données. De ce fait, les systèmes relationnels réagissent lentement aux modifications des exigences en matière de données. Les données non structurées telles que les fichiers binaires, audio et images ou les données semi-structurées telles que JavaScript Object Notation (JSON) doivent être stockées dans des systèmes non relationnels, communément No-SQL[12]. Dans ce paradigme, la structure de données n'est pas définie au moment de la conception et les données sont généralement chargées dans ces systèmes dans leur format brut. La structure des données n'est définie que lors de la lecture des données.

[12] No-SQL désigne une famille de systèmes de gestion de base de données qui s'écarte du paradigme classique des bases relationnelles et de leurs propriété ACID (Atomicité, Cohérence, Isolation et Durabilité).

Il existe principalement quatre types de bases de données No-SQL :

- Bases clé/valeur : Utilisées pour stocker des informations sous forme d'un couple clé valeur.
- Bases orientées documents : Utilisées pour stocker des documents et leurs métadonnées pour faciliter la recherche.
- Bases orientées graphes : Utilisées pour stocker les données sous forme de nœud et de relation.
- Bases orientées colonnes : Utilisées pour stocker les données sous forme de colonnes.

ETL vs ELT : Le processus ETL (Extract, Transform, Load) extrait dans un premier temps les données brutes. Dans un second temps les données sont transformées pour être enfin chargées dans l'entrepôt de données. Un inconvénient de l'approche ETL est que l'étape de transformation peut prendre beaucoup de temps et bloquer les ressources du système source. Avec l'approche ELT (Extract, Load, Transform), les données sont immédiatement extraites et chargées dans un référentiel de données pour être transformées. Si la finalité entre les procédés ETL et ELT reste la même, la manière d'arriver au résultat diffère. Citons quelques bénéfices de l'ELT :

- Réduction des délais de chargement,
- Performance de traitement de gros volumes de données,
- Flexibilité pour prendre en charge les transformations, si la manière dont le service marketing doit transformer les données diffère de la manière dont le service des opérations requiert ces mêmes données.

Comme évoqué, la sophistication des données prend un rôle de plus en plus important dans toutes les industries et son utilisation représente une transformation nécessaire.

Analyse avancée, intelligence artificielle, Machine Learning : L'analyse avancée inclut des outils tels que la modélisation prédictive et l'apprentissage automatique. Ces outils peuvent notamment aider les utilisateurs à trouver des patterns dans les données ; par exemple, l'analyse prédictive peut identifier les clients susceptibles de changer de service (Churn Detection[13]).

Traditionnellement, la manière dont les systèmes consomment et activent les données se présente sous la forme de règles métiers. Dans une règle de gestion, la décision, les conditions et les actions exécutées sont préprogrammées dans le système. Au regard de l'explosion des règles de gestions et du temps de traitement de ces règles, il est aujourd'hui préférable d'avoir recours aux prédictions. De plus en plus d'algorithmes produisent des décisions automatisées qui simplifient l'interaction avec les employés, les fournisseurs ou les clients. Les technologies autour de la science des données sont variées, mais le **R** et plus récemment le **Python** tendent à être des références.

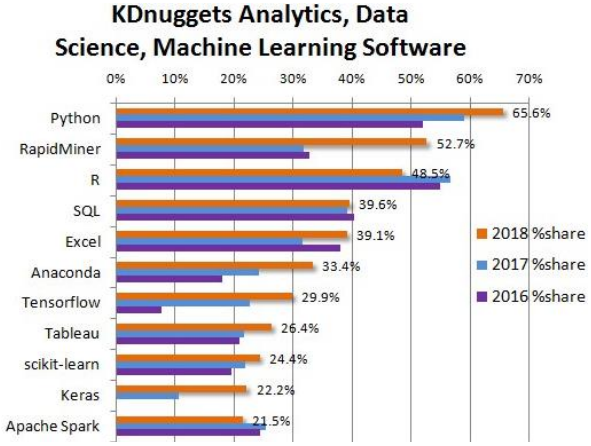


Figure 11 - Les technologies les plus utilisées en Machine Learning [16]

[13] La détection de l'attrition ou Churn est un processus de marketing prédictif permettant de détecter les clients risquant de quitter l'entreprise ou un service.

Pour répondre à l'indispensable besoin d'agilité, les utilisateurs doivent avoir accès à des outils leur permettant d'être autonomes dans l'analyses de données.

Self-Service : Les technologies dites de Self-Service Analytics rendent les données accessibles aux utilisateurs métier et favorisent ainsi la prise de décision rapide. Les utilisateurs finaux sont ainsi habilités et encouragés à effectuer des requêtes et à générer des rapports par eux-mêmes avec un support informatique limité. Les analyses en libre-service sont souvent caractérisées par des outils simples à utiliser dotés de capacités analytiques (mise à disposition de modèle pré-entrainé pour effectuer une estimation prédictive, un Forecast par exemple) et permettant de corréler différentes sources de données structurées ou non. Les solutions d'analyse en mode Self-Service sont également utiles pour tester à petite échelle des prototypes (POC : Proof Of Concept) et des modèles analytiques avant leur déploiement à plus grande échelle. Elles offrent généralement des bibliothèques de représentations visuelles telles que des graphiques, des cartes thermiques, des diagrammes de dispersion et de nombreuses solutions permettant aux utilisateurs d'élargir leurs options en important des visualisations à partir de sources extérieures, notamment open source.

Regardons comment ces différentes technologies peuvent s'articuler dans une architecture d'un projet décisionnel moderne :

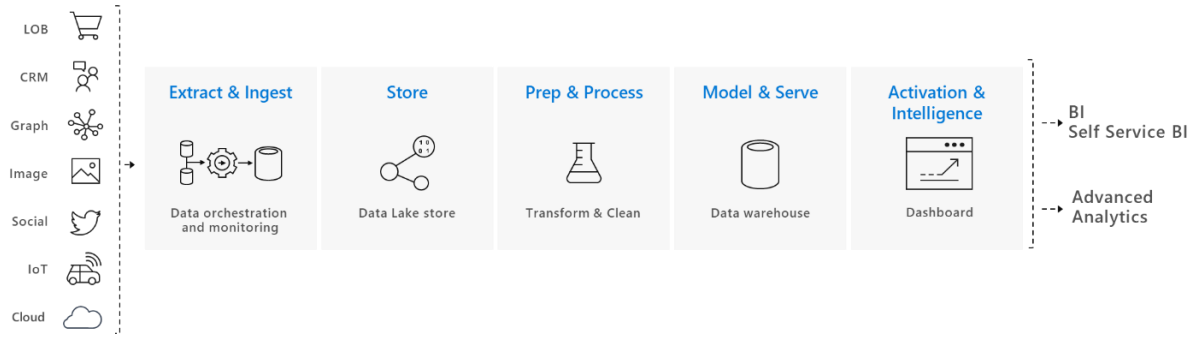


Figure 12 - { **Architecture V2** } - Système d'information décisionnelle moderne

Dans cette architecture, les données structurées et non structurées sont chargées en mode ELT dans un Data Lake. Ces données sont traitées par un cluster Big Data, des algorithmes de Machine Learning enrichissent l'information avant d'alimenter un Datawarehouse exposant les informations de manière performante facilitant ainsi la prise de décision.

Mobilité : Alors que le monde du travail se tenait dans un espace clos, aujourd'hui avec le numérique il est désormais possible de travailler partout et à tout moment. L'utilisation croissante des périphériques mobiles dans les entreprises en fait un contributeur majeur pour l'aide à la prise de décision rapide. Les outils d'analyse et de Reporting doivent prendre en compte ces nouveaux usages et modes de consommations de la donnée tout en adressant les nouvelles problématiques qui en découlent.

La BI mobile offre une grande disponibilité des informations (n'importe où et n'importe quand) et permet ainsi une vitesse de réaction plus rapide. Grâce à la fourniture d'applications mobiles appropriées à tous les utilisateurs d'appareils mobiles, les informations peuvent être utilisées par des personnes qui n'utilisaient pas auparavant les systèmes de BI traditionnelle. Cela conduit à un taux de pénétration de la BI supérieur dans les entreprises.

Mais nonobstant tous ces avantages, les entreprises sont confrontées à un certain nombre de défis lorsqu'elles tentent de réussir l'adoption d'une solution mobile [17] :

- Les entreprises et les développeurs d'applications doivent veiller à concevoir l'interface utilisateur de manière à garantir l'acceptation de l'application, en particulier dans les scénarii opérationnels où les travailleurs peuvent ne pas être habitués à ces outils.
- La sécurité et la confidentialité peuvent poser des problèmes. Les entreprises doivent veiller à mettre en place une configuration offrant une sécurité solide afin de protéger les données sensibles. En outre, la stratégie mobile devrait être alignée sur les procédures de sécurité existantes.
- Les appareils mobiles peuvent facilement être piratés, perdus ou volés. L'utilisation de la BI mobile peut par conséquent accroître les risques de violation des informations sensibles ou confidentielles.
- En raison de la taille réduite des écrans des périphériques mobiles, la conception d'applications de BI mobile présente de nouveaux défis pour les développeurs. Chaque appareil et navigateur fonctionne différemment.

La valeur des données se dépréciant avec le temps, il est nécessaire d'analyser celle-ci de plus en plus rapidement.

Temps réel : La génération de rapports et l'analyse rapide de données constituent un défi pour de nombreuses entreprises. En effet, il est de plus en plus nécessaire de rendre immédiatement disponibles les données des systèmes transactionnels afin de permettre une prise de décision opérationnelle plus rapide et factuelle. L'analyse en temps réel consiste à capturer des évènements ou des données immédiatement après leur apparition et à les traiter pour les afficher ou pour les analyser. Le volume de données en constante augmentation et la reconnaissance des formes d'évènements (traitement d'évènements complexes) ne sont que quelques-uns des défis. À l'instar de l'analyse visuelle ou de l'analyse prédictive, l'analyse en temps réel peut compléter la stratégie existante d'une organisation autour de la donnée afin d'obtenir de nouvelles informations (Référence au chapitre 2.2).

Les technologies de **Complex Event Processing** (CEP) permettent de traiter des évènements en continu et d'en extraire des informations pertinentes. Le CEP emploie des techniques telles que la détection de corrélation entre les évènements, les liens de causalité, l'analyse des évènements au regard de la chronologie, …

L'architecture **Lambda** permet de traiter d'énormes quantités de données en tirant parti des méthodes de traitement par lots (Batch Layer), des traitements de flux de type Stream (Speed Layer) et d'une couche de mise à disposition des données (Serving Layer). Ainsi cette architecture générique est en mesure de permettre l'analyse des données en temps réel et de les corréler avec les données historiques rendues intelligibles.

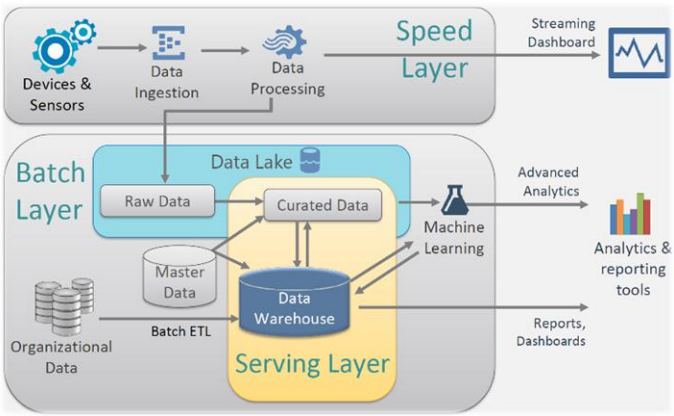


Figure 13 - Architecture Lambda [18]

L'architecture **Kappa** a été pensée pour pallier la complexité de l'architecture Lambda. Elle repose sur le principe de fusion de la couche temps réel et batch, ce qui la rend moins complexe que l'architecture Lambda

Actualisons l'architecture précédente pour y ajouter les technologies d'analyse en temps réel :

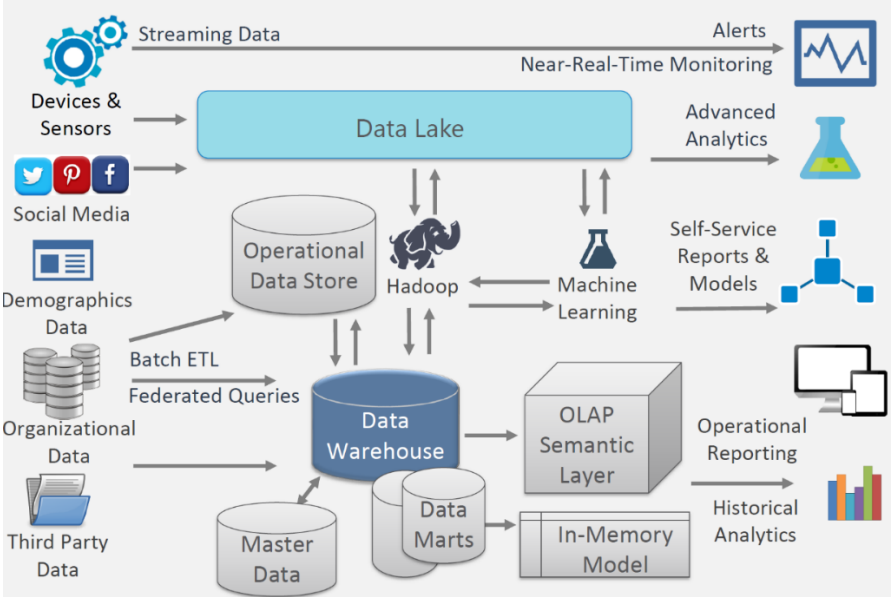
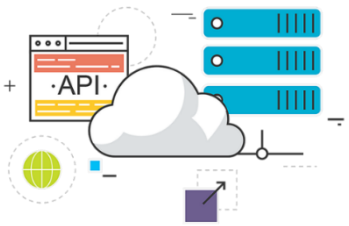


Figure 14 – { **Architecture V3** } - Moderne Datawarehouse

Dans un contexte d'interconnectivité en expansion constante où il est nécessaire d'obtenir une vision complète au-delà des frontières organisationnelles et technologiques afin d'améliorer l'efficacité de l'entreprise, il est nécessaire que les différents acteurs de l'EIM[14] puissent s'échanger des informations et communiquer entre eux. Les interfaces de programmation d'applications (API) permettent ces interactions.



API Economy : Les interfaces de programmation d'applications permettent l'accès aux services et données d'un tiers. À titre d'exemple les APIs de type Maps permettent d'intégrer à un site internet un service de cartographie, permettant de se focaliser uniquement sur le cœur de métier. Ainsi les ressources d'une organisation peuvent être réutilisées, partagées et monétisées via des API qui peuvent étendre la portée des services existants ou générer de nouveaux flux de revenus.

Les API peuvent ainsi se retrouver soit en amont de de l'architecture Data comme étant une source de données soit en aval comme étant une source de consommation des données rendue intelligible. D'autre part certaines API peuvent fournir de l'enrichissement de données et ainsi plutôt que de développer un modèle de calcul de sentiment de texte, il est possible d'appeler une API en lui passant en paramètre le texte à qualifier. Exemple avec les Cognitive Services d'Azure : Text Analytics.

[14] La gestion de l'information d'entreprise (EIM) est une discipline qui permet de structurer, décrire et gérer les actifs informationnels afin d'améliorer l'efficacité et de promouvoir la transparence des données au sein des entreprises.

Dans un souci d'agilité, les technologies de virtualisation et de conteneurisation se sont imposées rapidement dans le monde de l'informatique. Ils permettent de séparer une charge de travail du matériel sous-jacent.

Cloud Computing : D'un point vu financier, ce modèle offre des avantages indéniables. Il fournit aussi la flexibilité technologique nécessaire pour tirer parti des données en tant qu'actif stratégique. Le Cloud Computing permet d'exploiter la puissance de calcul ou de stockage de serveurs informatiques distants par l'intermédiaire du réseau. Les avantages du Cloud ne sont plus à démontrer : gestion simplifiée, flexibilité, évolutivité, réduction des coûts, CAPEX (Dépenses d'investissements pour l'achat du matériel) vs OPEX (Dépenses inhérentes à l'activité elle-même, dépense en fonction du besoin) …

Les fournisseurs de Cloud offrent trois services principaux :

- « Infrastructure as a Service » (IAAS) : Il s'agit d'ouvrir l'accès à un parc informatique virtualisé, des machines virtuelles sur lesquelles le consommateur peut installer un système d'exploitation et des applications.
- « Platform as a Service » (PAAS) : Cela consiste à offrir des ressources telles que le stockage, des ressources d'exécution, bases de données, etc. La responsabilité des tâches liées à la configuration et à la mise en œuvre dépend du consommateur.
- « Software as a Service » (SAAS) : Il s'agit d'offrir des applications, le consommateur n'a pas à se soucier d'effectuer des mises à jour, d'ajouter des patches de sécurité et d'assurer la disponibilité du service.

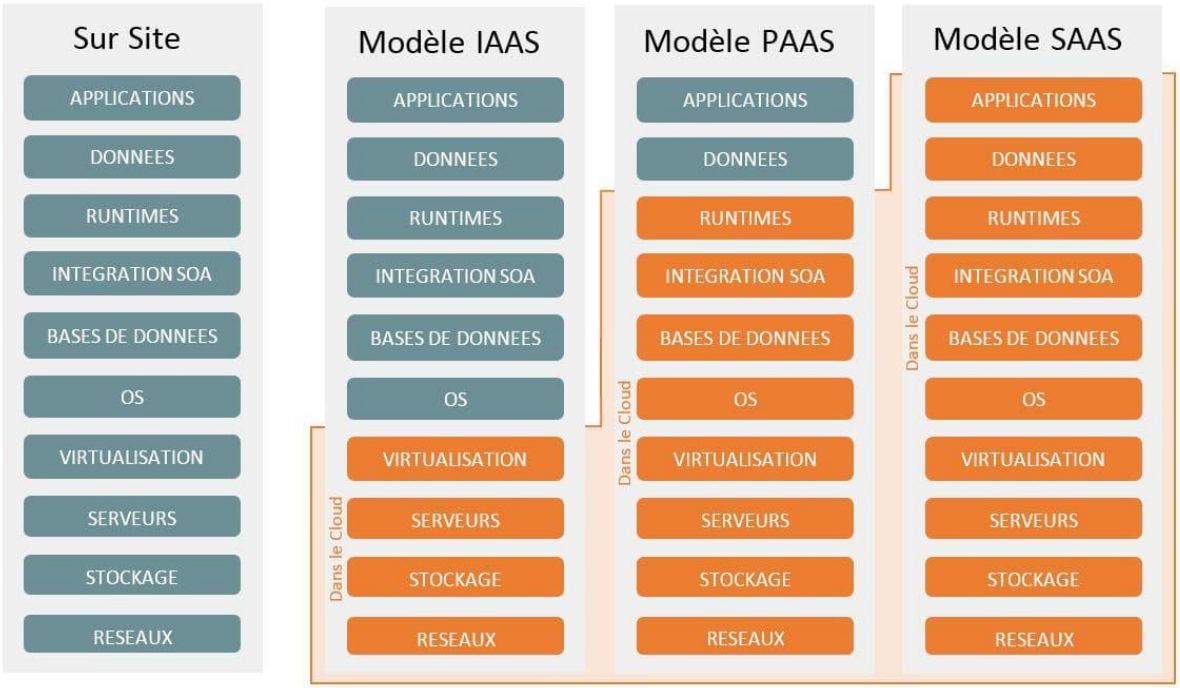


Figure 15 - Modèles de services de Cloud Computing

Les plateformes de traitement de la donnée dans le Cloud sont différentes de celles conçues pour l'infrastructure sur site (On-Premise). En effet la séparation entre le calcul et le stockage est une propriété architecturale importante du Cloud offrant une meilleure agilité. Afin de profiter au mieux des avantages du Cloud nous conseillons l'usage du SAAS plutôt que celui du au IAAS.

Exemple d'architecture IoT[15] en mode SAAS dans différents Cloud Provider :

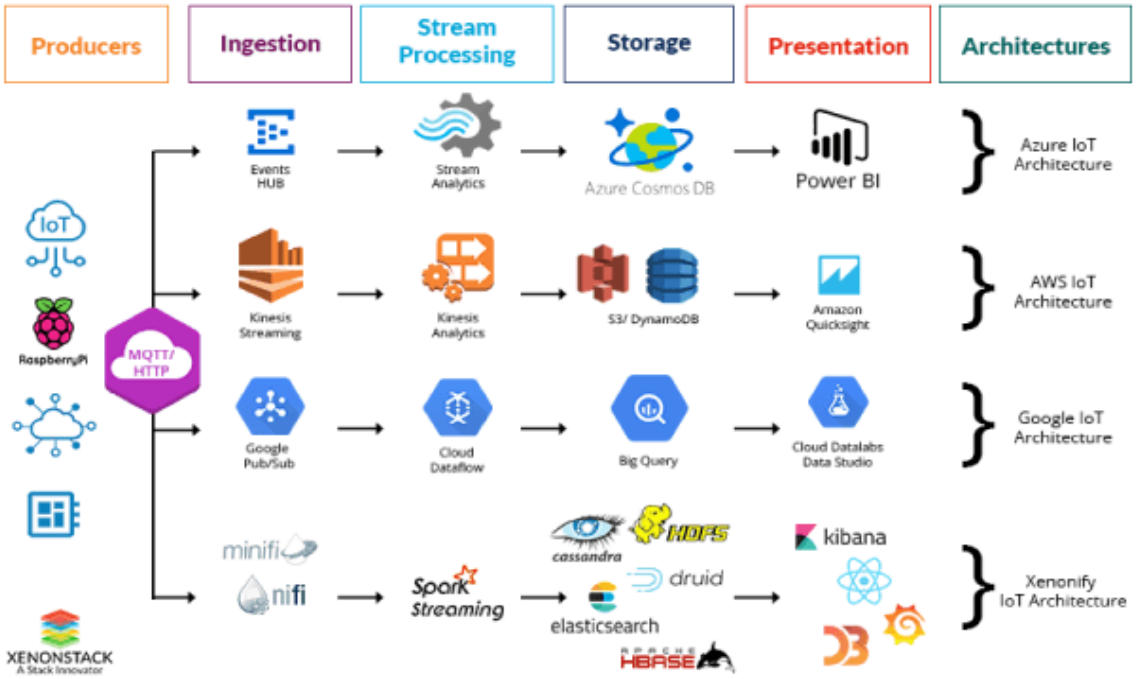


Figure 16 - IoT Cloud Architecture (Source : Xenonstackv)

Services AWS :

- Kinesis Streaming : Service de diffusion de données en temps réel.
- Kinesis Analytics : Service de traitement des données de streaming.
- S3 : Amazon Simple Storage Service est un service de stockage d'objet.
- DynamoDB : Service de base de données de clés-valeurs et de documents.
- QuickSight : Service rapide d'aide à la décision.

Services Google :

- Cloud Pub/Sub : Système de messagerie capable de gérer un flux de données en temps réel.
- Cloud Dataflow : Service de traitement des données par flux et par lots.
- Big Query : Entrepôt de données.
- Cloud Datalabes : Un outil interactif pour le Machine Learning et l'exploration, l'analyse et la visualisation de données.

Services Xenonstack :

- Apache NiFi : Logiciel libre de gestion de flux de données. Il permet de gérer et d'automatiser des flux de données entre plusieurs systèmes informatiques, à partir d'une interface web et dans un environnement distribué.
- Spark Streaming : Spark Streaming facilite la création d'applications de diffusion en continu évolutives et tolérantes aux pannes.
- Elasticsearch : Moteur de recherche et d'analyse distribué.
- Druid : Base de données d'analyse en temps réel hautes performances.
- Kibana : Kibana permet de visualiser des données présentes dans Elasticsearch.
- Grafana : Outil de datavisualisation.
- D3.JS : Bibliothèque JavaScript qui permet l'affichage de données sous une forme graphique.

[15] L'Internet des objets, ou IoT (Internet of Things) fait références à l'interconnexion entre Internet et des objets.

Services Azure (Cloud Microsoft) et actualisons de l'architecture précédentes :

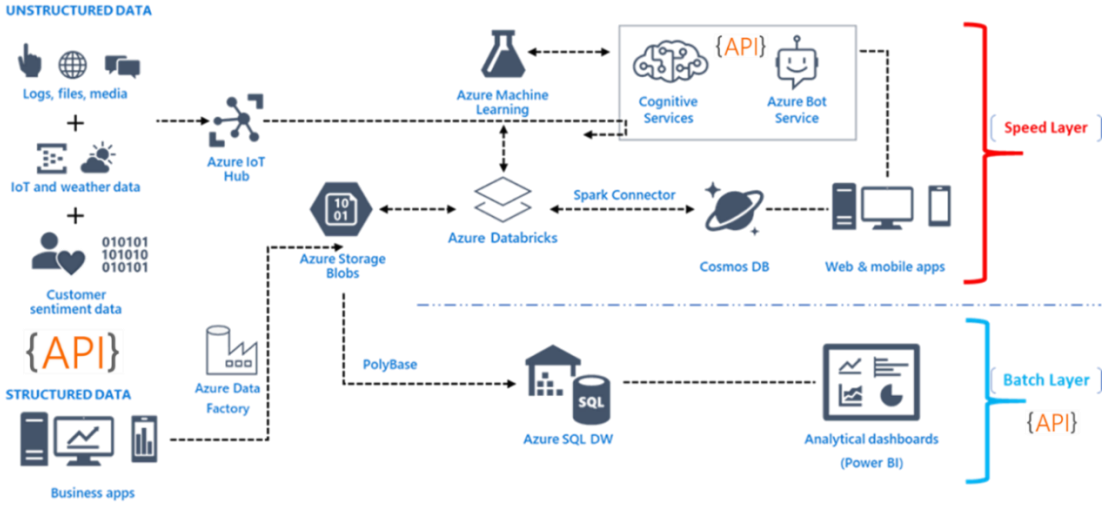


Figure 17 – { **Architecture V4** } - Analyse avancée dans Azure

- Event Hub : Service d'ingestion de données en temps réel.
- Stream Analytics : Service de traitement d'évènements complexes (CEP) en continu.
- Cosmos DB : Base de données No-SQL multi modèle distribuée à l'échelle mondiale.
- Power BI : Outils décisionnels de visualisation interactive des données.
- Azure IoT Hub : Service de concentrateur de messages (Message Queuing) permettant une communication sécurisée et bidirectionnelle entre des objets connectés et le Cloud.
- Azure Data Factory : Service d'orchestration des données.
- Azure Storage Blobs : Espace de stockage d'objets optimisé pour stocker de grandes quantités de données non structurées.
- Machine Learning Studio : Service collaboratif permettant de générer, tester et déployer des solutions d'analytique prédictive à partir de données.
- Cognitive Services : Il permet d'intégrer dans vos applications, sites web et Bots[16] des algorithmes intelligents pour voir, écouter, énoncer, comprendre et interpréter les besoins de vos utilisateurs au moyen de méthodes naturelles de communication.
- Azure Bot Services : Service de conception de Bots.
- Azure Databricks : Plate-forme d'analyse basée sur Apache Spark optimisée pour Azure.
- Azure SQL Datawarehouse : EDW (Enterprise Data Warehouse) basé sur une architecture MPP (Massively Parallel Processing[17]) pour exécuter rapidement des requêtes complexes sur des pétaoctets de données. La technologie Polybase permet de requêter des données non structurées en Transact-SQL sur un cluster Hadoop ou un Blob Storage.

Dans cette architecture, les données sont alimentées :

- En temps réel dans un service de queue d'évènements (**Speed Layer**), ces données sont ensuite enrichies par des modèles prédictifs puis sauvegardés dans une base de données No-SQL permettant leur analyse rapide. Ces données viendront ensuite enrichir un Datawarehouse.
- Dans un Data Lake (**Batch Layer**), une fois enrichies, elles alimentent le Datawarehouse. Enfin des Dashboard interactifs exposent les données et facilitent l'analyse.

[16] Un bot informatique est un agent logiciel qui interagit avec des serveurs, il offre une interface facilitant le dialogue entre un service et un consommateur. Le terme « bot » est la contraction par aphérèse de « robot ».
[17] En informatique, le traitement massivement parallèle est l'utilisation d'un grand nombre de processeurs (ou d'ordinateurs distincts) pour effectuer un ensemble de calculs coordonnés en parallèle.

Citons quelques avantages liés aux technologies Cloud par rapport aux traitements et aux stockages des données :

- Les données sont le moteur de la création de valeur, mais l'activation d'énormes quantités de données provenant de multiples sources ne peut se faire économiquement que dans le Cloud. En effet l'infrastructure étant flexible, elle permet de réduire les coûts et optimiser les performances de traitements (Scalability). Alors qu'une société aurait du mal à financer un cluster Hadoop de 50 nœuds, louer un tel cluster pendant une période courte à un Cloud Provider afin de tester la pertinence d'un projet est un atout indéniable.
- Les données sont de plus en plus générées dans le Cloud, il est donc logique d'analyser et de les traiter dans le Cloud.
- L'automatisation et le déploiement automatique de services sont facilités dans le Cloud.

Industrialisation : Dans un même souci d'efficacité et de réactivité, l'automatisation des tâches récurrentes est nécessaire. L'intégration continue et le déploiement continu (CI : Continuous Integration / CD : Continuous Deployment) sont des pratiques permettant d'itérer des modifications sur un projet rapidement tout en maintenant sa stabilité, les performances et la sécurité. Ainsi l'approche CI/CD garantit une automatisation et une surveillance continues tout au long du cycle de vie des applications, des phases d'intégration et de test jusqu'à la distribution et au déploiement. Les méthodes CI/CD s'appliquent aussi bien aux projets de développement applicatif qu'aux projets liés aux données.

Voici les différentes étapes nécessaires à la mise en place d'une Pipeline CI/CD :

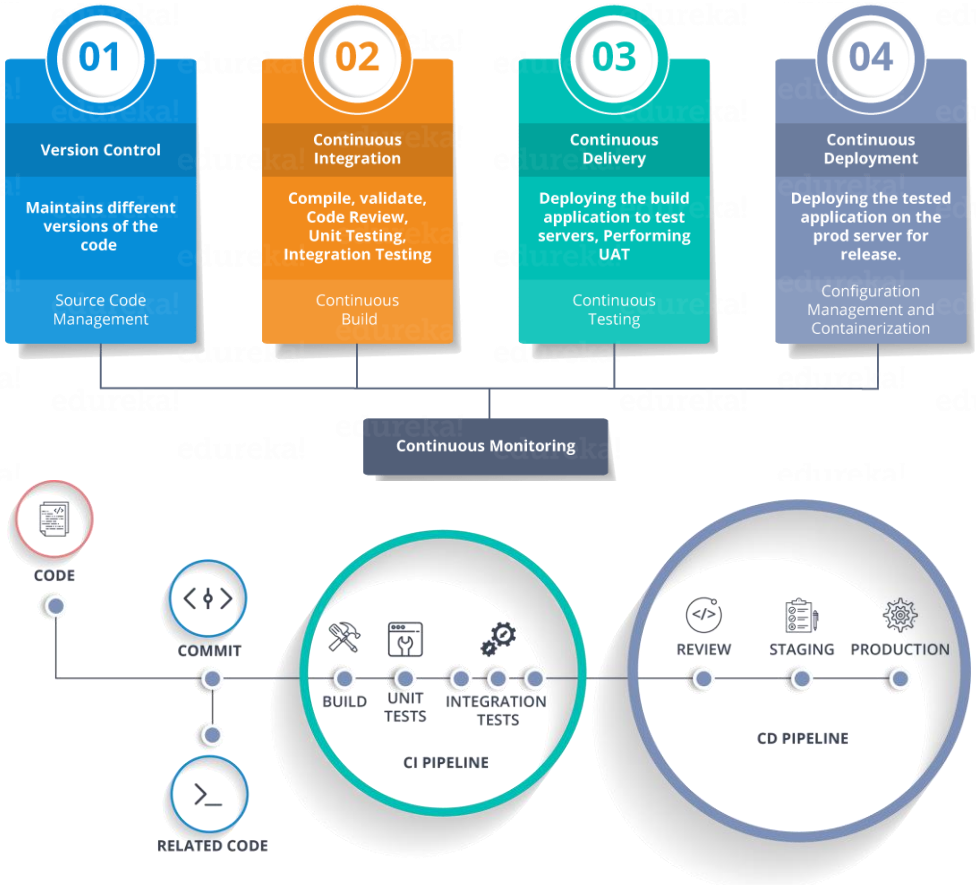
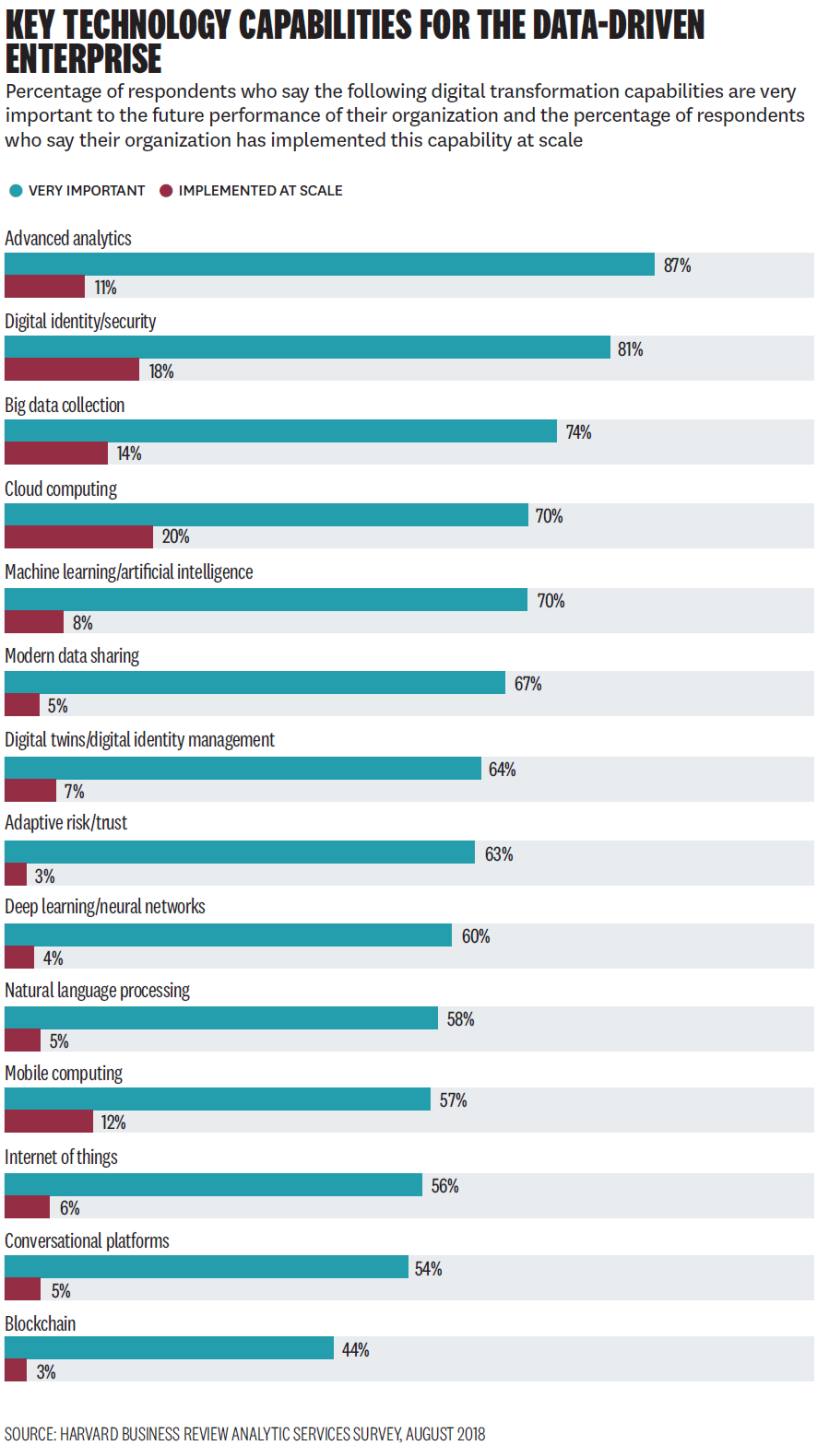


Figure 18 - CI/CD Pipeline [19]

Nous avons vu quelles sont les transformations technologiques nécessaires pour que les entreprises tirent parti des données. Analysons maintenant le point de vue de certaines entreprises, dans une étude menée par Harvard Business Review Analytic Services en août 2018 auprès de 729 dirigeants d'entreprise [20], bien que l'analyse avancée ne soit pas encore implémentée largement, elle représente la technologie clé pour les entreprises pilotées par les données. L'identité numérique, la sécurité, le Big Data, le Cloud et l'intelligence artificielle arrivent eux aussi en tête de la liste des fonctionnalités que les personnes interrogées jugent essentielles aux performances futures de leur entreprise.



b. Organisation

Utiliser les données efficacement ne passe pas uniquement par la technologie. En réalité, il s'agit d'une interaction complexe entre les données disponibles, les moyens et modes de stockage et la manière de les utiliser. Alors que la plupart des organisations se concentrent sur la technologie, les entreprises Leaders de leur secteur reconnaissent que les individus sont au centre de cette complexité. Par conséquent, bien que l'information et la technologie soient très pertinentes, il ne faut pas perdre de vue la variable humaine.

Embaucher et former ses collaborateurs : Il est important que tous les membres de l'organisation sachent utiliser les données pour prendre des décisions. Cela va à l'encontre de l'acronyme HiPPO[18] "Highest Paid Person's Opinion", utilisé pour décrire la tendance des employés les moins rémunérés à s'en remettre aux employés les mieux rémunérés lorsqu'une décision doit être prise. Ce terme peut également être utilisé pour décrire le fait qu'une organisation s'appuie sur son instinct humain plutôt que sur des données dans le processus de prise de décision.

Ouverture et implication : Les différentes entités de l'entreprise doivent avoir accès en mode self-service aux données et avoir été formées pour les utiliser et les interpréter afin d'en tirer pleinement parti. Cela peut passer par la formation d'utilisateurs clés (Key User). Par ailleurs, tous les employés des centres d'appels jusqu'aux représentants des ventes, doivent comprendre la valeur des données et leur rôle pour les rendre utilisables et être sensibilisés sur la qualité des données dès leur génération, ce qui évitera bon nombre de traitements en raison d'une qualité médiocre des données.

Afin d'encadrer ces transformations liées aux données, de nouveaux rôles sont apparus :

- Chief Data Officer (CDO), Head of Data : Ils ont à leur charge la stratégie Data de l'entreprise, la création d'un pôle Data ou d'un laboratoire de la donnée (DataLab) et le recrutement d'une équipe allant de l'Architecte à l'analyste en passant par les Data Engineer et Data Scientist.
- Data Engineer : Les Data Engineer ont pour rôles principaux de gérer les problématiques d'architectures, maintenir l'infrastructure, gérer les bases de données, concevoir des workflows, mettre en place une plateforme Big Data, du streaming, implémenter les algorithmes, assurer la scalabilité, la sécurité et la fiabilité des données.
- Data Scientist : Plus mathématiciens qu'informaticiens, les Data Scientist ont plusieurs domaines de compétence comme l'intelligence artificielle, le Machine et le Deep Learning, traitement automatique du langage naturel (NLP en anglais) et/ou développement d'algorithmes classiques.
- Ingénieur BI : Les développeurs BI ont une expertise sur la mise en place de jobs ETL, développement de Datamart et Datawarehouse ainsi que sur l'alimentation des différentes bases de données.

[18] HiPPO terme inventé par Avinash Kaushik dans le livre « Antithesis of data-drivenness ».

- Data Analyste, Data Steward : Le Data Analyste est le garant des données d'une entreprise. Sa mission est de permettre aux entreprises de comprendre et d'accéder à ces données provenant de sources diverses. Son rôle consiste également à recueillir, analyser, interpréter et lier les données provenant de différentes sources. Cela lui permet de réaliser des recommandations en les traduisant en actions à mener et objectifs business à atteindre. Le Data Analyste s'appuie sur de nombreux outils d'analyse, de reporting, de visualisation afin de les vulgariser et de les rendre plus compréhensibles (Data Storytelling[19]).

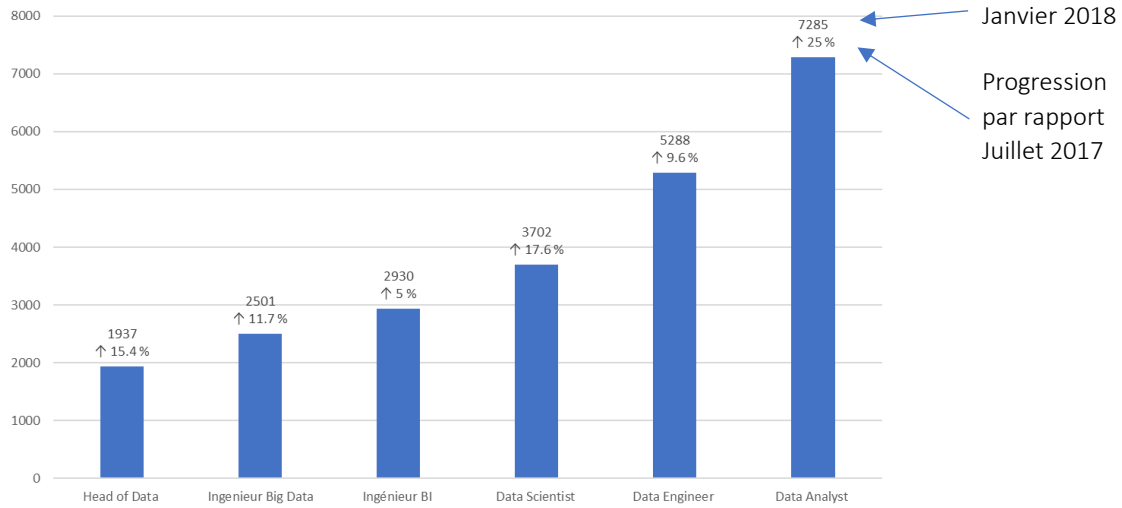


Figure 19 - Analyse de l'évolution de rôles autour de la donnée en France sur LinkedIn (Juillet 2017 et Janvier 2018) [21]

Ces différents rôles peuvent être réunis dans une entité organisationnelle dont la mission est de permettre à d'améliorer la performance de l'entreprise par l'exploitation de la Data : le **DataLab**. Il peut avoir la charge d'assurer l'identification, le développement et l'accompagnement à la généralisation des projets « Test & Learn »[20] autour de la Data.

Exemple d'organisation d'un DataLab :

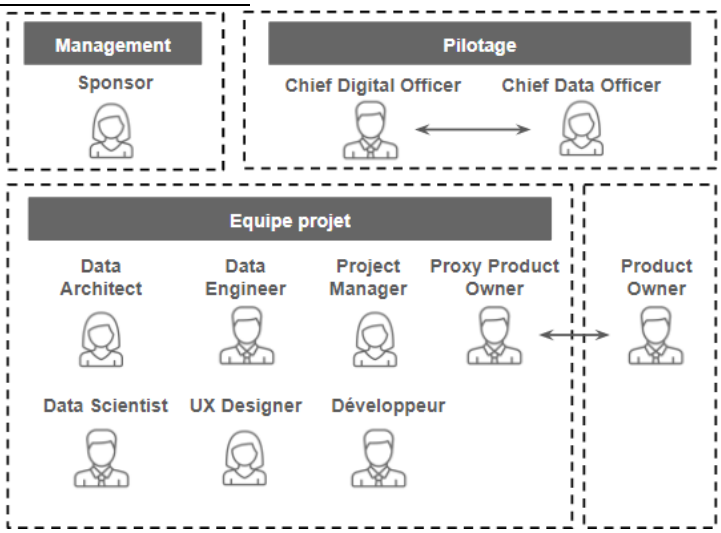


Figure 20 - Organisation dans un DataLab [22]

[19] Le Data Storytelling est la capacité de raconter une histoire avec les données, et de personnaliser les données en fonction de l'audience.

[20] La méthode de gestion de projet dite de "Test & Learn" consiste à mettre en place des projets sous-jacents au projet global et de mesurer rapidement leur efficacité afin d'évaluer la pertinence du projet.

Collaboration : Les employés sont de plus en plus confrontés à des problématiques singulières qu'ils rencontrent pour la première fois. Leur force est alors de trouver les meilleures réponses ou solutions à partir des informations disponibles. De ce fait, les entreprises doivent embaucher des personnes ayant des compétences en résolution de problèmes. Par ailleurs, les problèmes sont plus faciles à résoudre à plusieurs : **La capacité d'innover et de répondre à des problèmes dans une organisation est d'autant plus grande qu'il existe de liens et de collaborations entre les personnes.**

Le meilleur moyen de générer plus d'idées est de réunir les cerveaux de personnes différentes, ayant des connaissances différentes… Idéalement, lors de l'embauche, les managers doivent penser aux trois points suivants :

- Individuel : Ce que cette personne va apporter individuellement (Connaissance, autonomie et dynamisme attendu).
- Équipe : Ce qu'elle va apporter à l'équipe (en comblant ou corrigeant les points faibles de l'équipe).
- Travail : Le profil de l'équipe est-il en adéquation avec le besoin initial ?

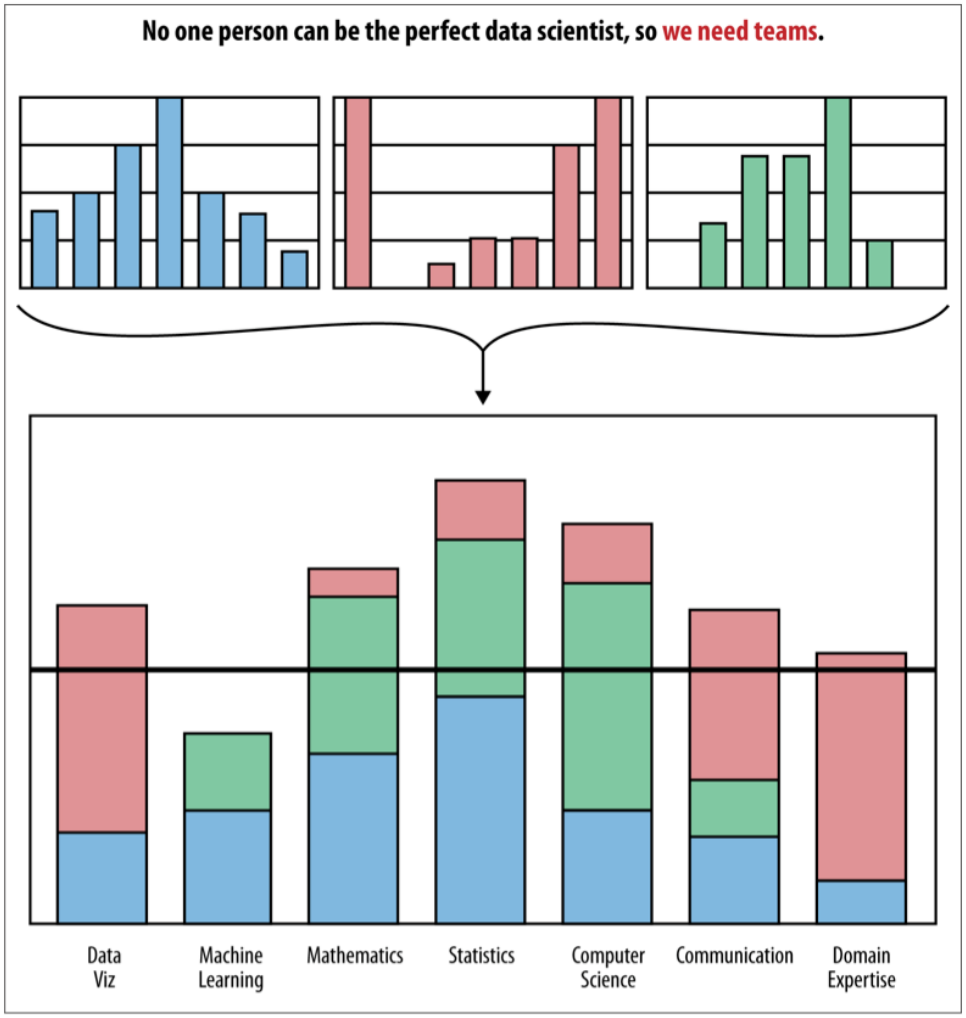
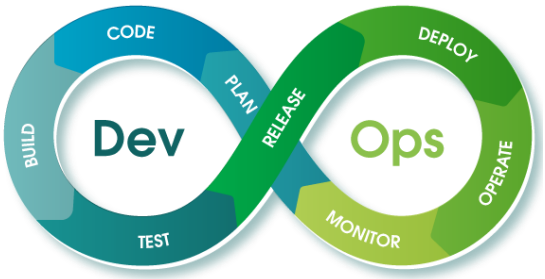


Figure 1-3. Data science team profiles can be constructed from data scientist profiles; there should be alignment between the data science team profile and the profile of the data problems they try to solve

Figure 21 - Data Science Team Profiles [23]

Agilité : Alors que les entreprises fonctionnent à un rythme de plus en plus rapide, si les données ne progressent pas au même rythme, elles sont retirées du processus de prise de décision. De même, alors que les données deviennent plus courantes et qu'elles constituent un avantage concurrentiel, le besoin de les démocratiser et de les rendre accessibles est ressenti très fortement dans les entreprises, et ce à tous les niveaux. Ceci est similaire à la manière dont l'agilité dans la création d'applications a conduit à la création de la culture **DevOps**[21]. La même agilité est maintenant également nécessaire du côté des données.



Les pratiques DevOps modifient la manière dont les applications sont développées et déployées au sein des entreprises. De même, le **DataOps** modifie la façon dont les données sont consommées. Voici la définition du Gartner [22] : "*Data ops is the hub for collecting and distributing data, with a mandate to provide controlled access to systems of record for customer and marketing performance data, while protecting privacy, usage restrictions and data integrity*".

Au même titre que les développeurs écrivent du code, les Data Scientist conçoivent des modèles analytiques pour extraire des informations exploitables à partir de gros volumes de données. Ils doivent avoir accès aux données, pour entrainer leurs modèles, et doivent pouvoir les déployer pour les mettre à dispositions des décideurs. Dans un tel environnement et avec une telle demande d'agilité, une culture DataOps devient une nécessité. Grâce au DataOps, l'entreprise est à même par exemple de mieux valoriser la science des données en réduisant les délais de mise en production. Le DataOps focalise ainsi le développement sur les données, au lieu de l'application elle-même. Comme le DevOps, le DataOps a son manifeste : http://dataopsmanifesto.org.

Le DataOps est avant tout une culture qui doit être supportée par des processus organisationnels, des actions et des bonnes pratiques sur le plan technologique. Elle tire parti du « DevOps » et de « Lean Manufactoring » pour harmoniser les interactions entre tous les utilisateurs de la donnée (Data Scientist, Data Analysts, Data Engineers, Développeurs, etc.) afin de réduire la durée du cycle de vie Data Analytics et augmenter la qualité de ses livrables.

Auparavant les informations devaient être surveillées de près et rendues accessibles uniquement aux personnes clés, les entreprises étant souvent incapables de concevoir que le partage de ces données dépassait l'intérêt de certains décideurs. Aujourd'hui, nous savons qu'il est extrêmement utile de partager des données plus largement. Cependant, même si le partage des données doit être encouragé, cela ne doit pas se faire au détriment de la gouvernance.

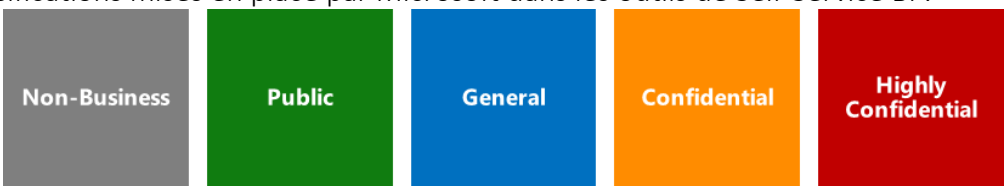
Gouvernance : Elle correspond à l'ensemble des procédures mises en place au sein d'une entreprise afin d'encadrer la collecte de données et leur utilisation. Il s'agit autant de respecter les obligations légales imposées par les pays que d'instaurer un cadre interne afin d'optimiser l'utilisation des données.

[21] Le DevOps est un mouvement en ingénierie informatique et une pratique technique visant à l'unification du développement logiciel (Dev) et de l'administration des infrastructures informatiques (Ops).
[22] Source : https://www.gartner.com/it-glossary/data-ops

La gouvernance fait aussi référence aux règles, politiques et procédures utilisées par les organisations pour décider quelles données doivent être conservées, pendant combien de temps, sous quel format, etc. **La gouvernance des données doit s'appliquer à l'entreprise et non à un sous-ensemble de celle-ci**.

Voici quelques aspects clés de la gouvernance de données :

- Politique de la donnée : Définir, approuver et faire respecter les règles de données définies par l'extérieur comme les règlementations (CNIL[23], HIPAA[24], GDPR[25], PCI DSS[26]) et les directives internes telles que la classification de données. À titre d'exemple, voici les classifications mises en place par Microsoft dans les outils de Self-Service BI :



- Surveillance et mesure : Intégration de méthodes de validation de la conformité aux stratégies de données définies et génération automatiquement d'alertes en cas de violation des stratégies de données.
- Qualité : Les données doivent être fiables avant d'être utilisées pour la prise de décision. Par conséquent, des processus et des outils doivent être mis en place en continu pour vérifier la cohérence et l'intégrité des données.
- Référencement des données dans un catalogue : Les métadonnées facilitent l'identification des données. Il est donc essentiel de gérer correctement les métadonnées pour assurer une bonne gouvernance des données à long terme. La création de métadonnées sur chaque actif de données facilite leur accessibilité.
- Protection des données : Mise en place de méthodes de cryptage et de masquage des données sensibles.

Pour une gouvernance efficace des données, les dirigeants doivent travailler en étroite collaboration avec le service informatique. Enfin, pour assurer la cohérence, les politiques mises en place doivent également être clairement communiquées à toutes les parties prenantes de l'organisation. Comme le soulignent Tom Davenport, Jeanne Harris et Robert Morrison dans leur livre Analytics at Work, « *Tout employé peut orienter une entreprise dans une direction plus analytique* » cependant le Sponsorship au niveau C-Level[27] est un facteur clé de réussite.

---

[23] La Commission Nationale de l'Informatique et des Libertés (CNIL) est une autorité administrative indépendante française chargée de veiller à ce que l'informatique soit au service du citoyen et qu'elle ne porte atteinte ni à l'identité humaine, ni aux droits de l'Homme, ni à la vie privée, ni aux libertés individuelles ou publiques.

[24] HIPAA, acronyme anglais de Health Insurance Portability and Accountability Act, est une loi votée par le Congrès des États-Unis en 1996 qui définit les normes américaines pour la gestion électronique de l'assurance maladie,

[25] GDPR, acronyme anglais de General Data Protection Regulation, est un règlement de l'Union européenne qui constitue le texte de référence en matière de protection des données à caractère personnel. Il renforce et unifie la protection des données pour les individus au sein de l'Union européenne.

[26] PCI DSS, acronyme anglais de Payment Card Industry Data Security Standard désigne les normes de sécurité des données applicables à l'industrie des cartes de paiement.

[27] C-Level, est un adjectif utilisé pour décrire les titres exécutifs de haut niveau au sein d'une entreprise. La lettre C, dans ce contexte, signifie chef. Les personnes qui occupent des postes de niveau C sont généralement considérés comme les membres les plus puissants et les plus influents d'une organisation : Chief Executive Officer (CEO), Chief Financial Officer (CFO), Chief Information Officer (CIO), Chief Technology Officer (CTO), …

Formation : Aujourd'hui, plus que jamais, il est nécessaire d'apprendre continuellement et c'est d'autant plus un enjeu stratégique que le rythme d'obsolescence des compétences métiers est passé en moyenne à 18 mois. Ce chiffre est le résultat d'une accélération incroyable des innovations technologiques. Dans ce contexte, les entreprises doivent faciliter l'apprentissage continu de leurs employées.

Veille technologique. Alors que les technologies et les règlementations évoluent de plus en plus rapidement, la mise en place d'une veille, qu'elle soit technologique, règlementaire ou stratégique, nécessite une certaine organisation et de bonnes pratiques. La veille informationnelle est une activité continue et en grande partie itérative permettant d'anticiper les évolutions. Elle correspond ainsi à une **nécessité pour toutes entreprises souhaitant tirer parti de ces données en tant qu'actif stratégique**. La démarche de veille peut s'organiser autour des étapes suivantes :

- La définition des axes de veille ;
- L'organisation de la veille ;
- L'identification et la définition des sources ;
- La démarche de collecte et de surveillance de l'information ;
- L'analyse, la synthèse et l'exploitation des résultats de veille, le partage et la diffusion.

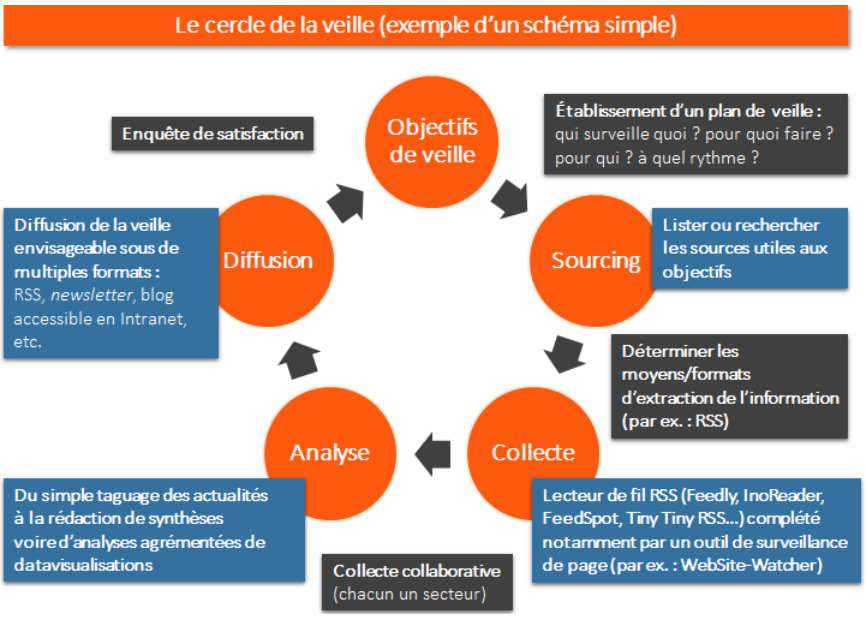


Figure 22 - La veille est un processus circulaire où chaque étape se doit d'être documentée [24]

La veille technologique doit être réalisée à tous les niveaux :

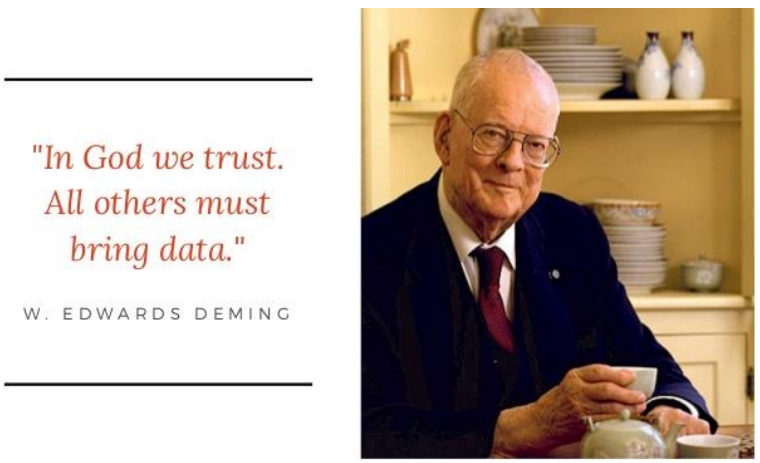
- Personnel : Chacun des employés doit mettre en place sa propre veille. En effet, d'après une étude de Dell et de l'Institut pour le Futur datant de juillet 2017, 85% des emplois en 2030 n'existent pas encore aujourd'hui [25].
- Projet : Afin d'anticiper les potentielles évolutions.
- Direction des Systèmes d'Information (DSI) : Pour offrir des solutions aux nouveaux besoins business.
- Entreprise : La rentabilité d'une structure ne se bâtit pas uniquement sur l'aspect opérationnel. Une entreprise a plus d'impact lorsqu'elle agit proactivement plutôt que de réagir à un évènement, car dans ce cas, elle a par définition toujours un temps de retard. Développer la veille en entreprise, c'est savoir identifier et analyser les informations du marché, prendre du recul pour ne pas être déstabilisé et ainsi être plus fort sur le terrain grâce une stratégie claire et efficace pérennisant son activité.

c. Culture

Utiliser les données efficacement ne passe pas uniquement par la technologie ou encore l'organisation. Cela implique un changement culturel[28] profond au sein de l'entreprise. La littératie des données (du terme anglais **Data Literacy**) ou culture des données est la capacité d'identifier, collecter, traiter, analyser, interpréter des données afin de comprendre les phénomènes, les processus, les comportements qui les ont générées.

> *"A data culture isn't just about deploying technology alone, it's about changing culture so that every organization, every team and every individual is empowered to do great things because of the data at their fingertips." Satya Nadella [26]*

Culture de la mesure : "*If You Can't Measure It, You Can't Improve It*", dicton difficilement réfutable de William Edwards Deming. Deming a été fortement impliqué dans la reconstruction économique du Japon après la Seconde Guerre mondiale, pays qui fut alors le moteur économique mondial des années 60, 70 et 80. Deming avait pour philosophie fondamentale que la mesure et l'analyse des données étaient essentielles pour obtenir des performances supérieures dans tous les domaines.



Cette seconde citation concise souligne l'importance de la donnée pour toute prise de décision. Ainsi, les entreprises du Web ont pour la plupart déjà pris l'habitude de tout mesurer en passant par le nombre de visites, les temps de réponse, les pages les plus vues, le temps passé sur chaque page… Mais en mesurant davantage comme la chaleur dégagée par les processeurs ou encore la consommation électrique des transformateurs, ils ont aussi pu faire des économies ... Cela permet d'ajuster et d'optimiser l'efficacité énergétique des installations informatiques (PUE[29]). Ils ont alors appris à baser leurs plans d'action sur cette masse d'informations.

Dans la continuité de la mesure constante, l'**A/B Testing** consiste à comparer deux versions d'une page Web ou d'une application afin de vérifier laquelle est la plus performante.

[28] La culture d'entreprise est un ensemble de connaissances, de valeurs et de comportements, partagés par la plupart de ses membres, qui facilitent le fonctionnement d'une entreprise.
[29] L'indicateur d'efficacité énergétique (en anglais PUE ou Power Usage Effectiveness) est utilisé pour qualifier l'efficacité énergétique d'un centre d'exploitation informatique. C'est un des éléments de l'informatique éco-responsable (Green IT).

Les variations présentées ci-dessous, dénommées A et B, sont présentées de manière aléatoire aux utilisateurs. Une partie d'entre eux sera alors dirigée vers la première version tandis que l'autre sera affectée à la seconde.

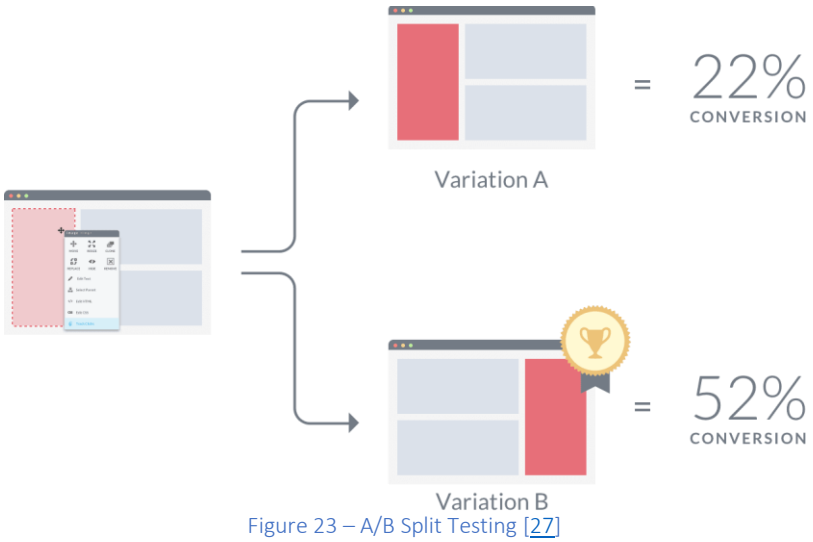


Figure 23 – A/B Split Testing [27]

Une analyse statistique permet par la suite de tester l'efficacité des versions A et B sur différents indicateurs comme le taux de conversion. En d'autres termes, on peut vérifier quelle version déclenche le plus de clics, d'abonnements, d'achats… Les résultats permettent alors de déterminer la meilleure stratégie à adopter.

Nous l'avons vu dans cet exemple issu du Web, la culture de la mesure c'est-à-dire le réflexe de collecter de la data par la métrologie[30] pour calculer des indicateurs permettant de prendre des décisions, s'invite aujourd'hui dans tous les domaines (IoT, Industries, …). De même le Lean Startup évoqué dans le chapitre 2.3 repose principalement sur les mesures pour itérer et améliorer produits ou services.

> *"Metrics influence the behavior of the entire organization and affect both actions and decisions related to future strategies", John Hauser [28]*

Ouverture : Les entreprises s'appuient beaucoup sur les données internes de l'entreprise pour prendre les décisions et sont plus frileuses face à l'exploitation de données externes. Or, le plus souvent, étudier et éventuellement intégrer les données externes permet d'explorer facilement et rapidement des périmètres parfois moins bien connus, mais qui pourtant peuvent être vecteurs de valeur.

Stratégie : Il est nécessaire pour les entreprises où l'analyse de données est une réalité omniprésente d'avoir une culture de la donnée. Plusieurs études menées par le cabinet McKinsey suggèrent que la culture autour de la donnée ne peut être imposée[31]. De ce fait une stratégie d'entreprise doit être créée.

[30] La métrologie est la science de la mesure. Elle définit les principes et les méthodes permettant de garantir et maintenir la confiance envers les mesures résultant des processus de mesure. Il s'agit d'une science transversale qui s'applique dans tous les domaines où des mesures quantitatives sont effectuées.
[31] Source McKinsey : Why data culture matters.

Dans l'implémentation de la stratégie autour de la donnée, l'engagement de la direction est essentiel, en effet les hauts responsables doivent promouvoir la transparence des données à tous les niveaux. Tout en suivant leurs politiques d'accès sécurité aux données, le Risk Management devrait fonctionner comme un accélérateur intelligent d'accès responsabilisé. Pour aller plus loin, et ainsi créer un avantage concurrentiel, il est nécessaire de stimuler la demande de données au plus bas niveau dans l'entreprise.

"The best advice I have for senior leaders trying to develop and implement a data culture is to stay very true to the business problem: What is it and how can you solve it? If you simply rely on having huge quantities of data in a data lake, you're kidding yourself. Volume is not a viable data strategy. The most important objective is to find those business problems and then dedicate your data-management efforts toward them. Solving business problems must be a part of your data strategy." Rob Casper, Chief Data Officer, JPMorgan Chase.

La Direction des Systèmes d'Information doit jouer le rôle d'accélérateur de transformation digitale et tenir son rôle de conseil sur la manière dont les technologies influent sur le Business. Elle doit par ailleurs fournir un socle numérique agile, robuste et sécurisé pour opérer une transformation digitale viable, anticiper l'évolution des besoins métiers, être force de proposition et suggérer de nouvelles solutions en phase avec les innovations numériques. Ainsi l'une des difficultés des DSI est de faire cohabiter « Core IT et Fast IT »[32]. La dynamique numérique constante nécessite parfois l'appel à l'externalisation de services, de données … La DSI doit aujourd'hui apprendre à composer avec les startups et les partenaires externes qui peuvent lui apporter un point de vue visionnaire et une source d'inspiration.

La Direction des Ressources Humaines (DRH), en charge du recrutement permettant de soutenir la culture digitale de l'entreprise doit prévoir l'obsolescence des compétences de ses collaborateurs et les accompagner dans une mise à niveau. Il ne faut donc pas la négliger mais au contraire lui accorder la position stratégique qui lui revient.

Adoption : L'utilisation des données appliquée à un problème métier crée de l'innovation. Les entreprises doivent fournir l'accès aux données à leurs collaborateurs et leur permettre de donner suite à leurs idées novatrices et ainsi créer de la valeur. Lorsque l'enthousiasme suscité par l'analyse des données imprègne toute l'organisation, il devient une source d'énergie et de dynamisme.

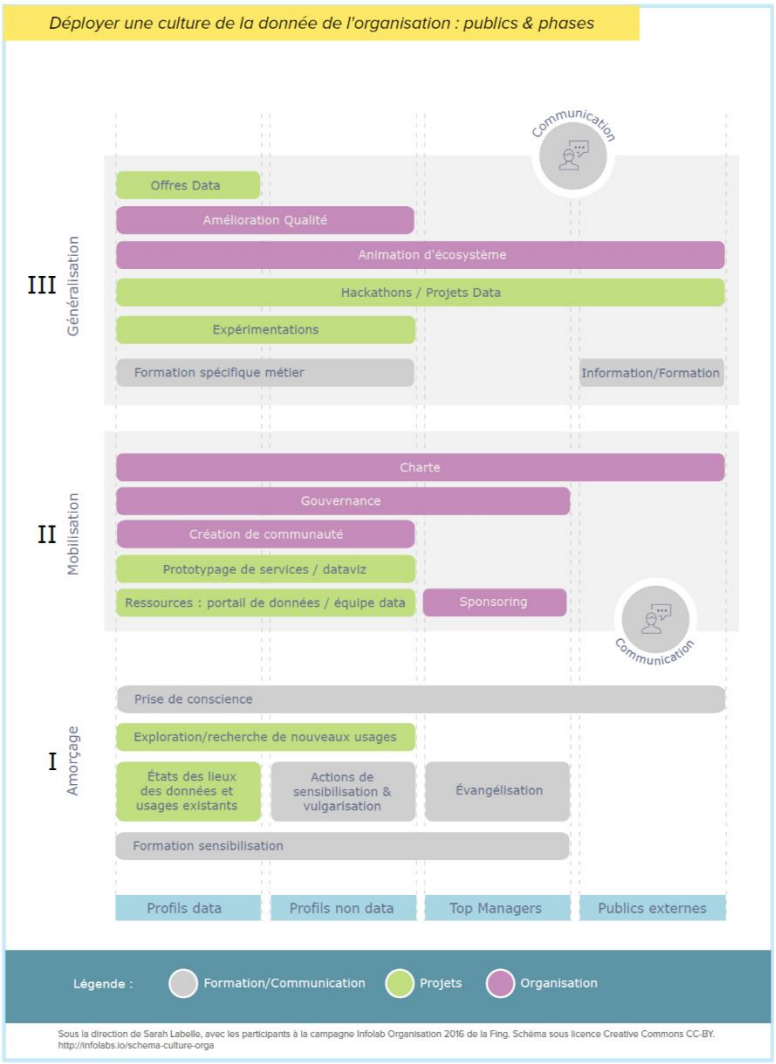
"You have to figure out how to really democratize the data-analytics capability, which means you have to have a platform through which people can easily access data. That helps people to believe in it and to deliver solutions that don't require an expensive data scientist. When people begin to believe in the data, it's a game changer: They begin to change their behaviors, based on a new understanding of all the richness trapped beneath the surface of our systems and processes." Ted Colbert, CIO, Boeing.

[32] Le terme de « Core IT » est utilisé parler des systèmes d'information hérités de toutes les évolutions qui ont eu lieu jusqu'à aujourd'hui, également appelé le SI existant ou le « Legacy ». Le terme « Fast IT » est quant à lui utilisé pour parler d'informatique agile utilisant des technologies innovantes, exploitant la donnée (produite, stockée, partagée, analysée) pour répondre à de nouveaux usages, cultures et organisations.

Éthique : Alors que l'intelligence artificielle offre de grandes opportunités pour prospérer ou accroitre sa croissance, celle-ci engendre également certains risques qu'il convient de gérer correctement. Il convient de trouver et maintenir un équilibre entre les avantages de l'innovation et les inconvénients des risques. L'éthique entre en jeu, en partie, lorsque les organisations réalisent que l'information a une valeur qui peut être extraite et transformée en de nouveaux produits et services. Les entreprises qui n'évaluent pas explicitement et de manière transparente les impacts éthiques des données qu'elles collectent auprès de leurs clients risquent de nuire à la qualité de leurs relations. L'évaluation éthique inclut à la fois une compréhension de la manière dont une organisation utilisera les données client et une compréhension des valeurs que l'organisation défend.

Le concept de "**Privacy by Design**" a pour objectif de garantir que la protection de la vie privée soit intégrée dans les nouvelles applications technologiques et commerciales dès leur conception. Pour chaque nouvelle application, produit ou service traitant des données à caractère personnel, les entreprises et autres responsables du traitement devraient offrir à leurs utilisateurs ou clients le plus haut niveau possible de protection des données.

Déploiement : Enfin, déployer la culture de la donnée dans l'organisation requiert plusieurs phases. La plus essentielle est de s'intéresser aux pratiques des collaborateurs pour comprendre la place que les données auront dans leur quotidien. Qu'ils en produisent ou en manipulent, les collaborateurs ne possèdent pas forcément une conception très structurée de leur rôle au sein de l'organisation. Selon les cas, le partage et la diffusion d'une culture des données vont permettre le développement d'une approche transversale et cohérente du sujet de la donnée au sein de l'entreprise. Ainsi Infolab a synthétisé cela dans le schéma suivant :



Derrière ces opportunités stratégiques, l'analyse des données bouleverse le fonctionnement et la culture des entreprises. Afin d'être en mesure de profiter de ces transformations liées à la donnée et en tirer parti en tant qu'actif stratégique, il faut pour l'entreprise :

- Utiliser les technologies adéquates, étant précisé que le Data Lake et les bases de données No-SQL offrent plus de flexibilités par rapport au SGBD traditionnel. Les analyses avancées offrent de nouvelles possibilités quant à l'analyse et l'extraction de valeurs des données brutes. La self-service BI et les outils mobiles offrent une agilité et une indépendance accrue des employées. L'architecture Lambda est souvent utilisée dans les architectures temps réel. Enfin, le Cloud permet une flexibilité aux équipes et facilite l'automatisation.
- Embaucher les ressources en adéquation avec les besoins de l'entreprise. Former ses employés, les impliquer et encourager leur collaboration, leur fournir l'agilité et la gouvernance nécessaire. Enfin, l'organisation en charge de la gouvernance des données doit non seulement faire respecter les règlementations internes et externes à l'entreprise, mais également maximiser le potentiel des données utilisées.
- Adopter une culture de la mesure et avoir une stratégie favorisant l'adoption et suscitant l'enthousiasme pour l'analyse des données dans toute l'organisation tout en respectant les valeurs de l'entreprise, son éthique (« Garde-fou à l'innovation »).

Les données sont un puissant vecteur de développement économique et de production de connaissances, mais elles suscitent également des inquiétudes légitimes, ainsi que des luttes de pouvoir. Ainsi ces transformations doivent être soutenues par la direction et les données doivent être traitées comme des actifs de l'entreprise. Les entreprises doivent fournir le financement, la dotation en personnel et l'attention qu'elle accorde aux autres actifs de son organisation. Pour résumer, voici les transformations nécessaires à toutes entreprises voulant tirer parti de ces données :

Data Company = Technologie + Organisation + Culture

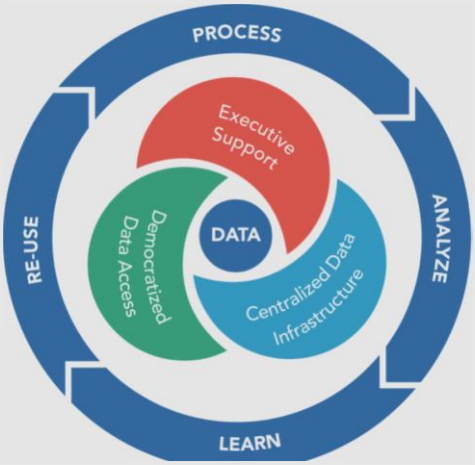
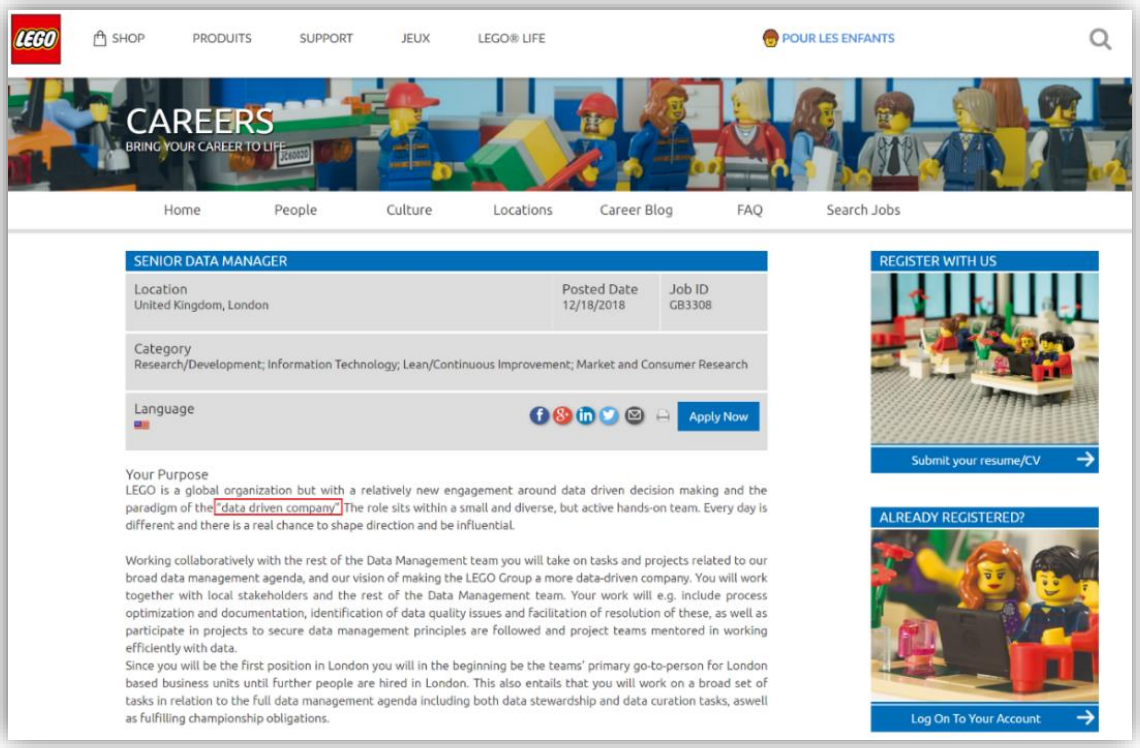


Figure 24 - The aspirations for a data-driven enterprise, source : Qubole.

3. Data Companies

Aujourd'hui, *Buzz Word*, on entend parler de Data Company, de Data Driven Company, de Data Driven business un peu partout jusqu'aux offres d'emploi pour des sociétés dans lesquelles les données ne sont clairement pas le cœur de métier. Exemple avec une offre d'emploi de la société de jouet Lego :



Est-ce un effet de mode, ou est-ce une nécessité dans le cadre de la transformation numérique des entreprises d'intégrer au centre de leur culture la donnée même si cela n'est pas le cœur de leur métier ?

Dans ce chapitre nous tâcherons d'apporter une définition précise de ce qu'est et de ce que n'est pas une Data Company.

3.1 – Définition

Une entreprise Data Driven est une entreprise « pilotée par les données ». Autrement dit, il s'agit d'une entreprise qui s'appuie sur l'analyse des données qu'elle a à sa disposition pour prendre des décisions et orienter son évolution.

Les organisations orientées données se développent à un rythme fulgurant et délogent des acteurs bien établis dans une grande variété de secteurs. Ainsi, Uber qui est la première société de taxi au monde ne possède pas de véhicules, Alibaba qui est Leader de la vente au détail n'a pas d'inventaire et Airbnb qui est Leader mondial des offres d'hébergement ne possède aucun bien immobilier. Leurs actifs physiques ne contribuent que de façon marginale à l'incroyable valorisation de ces jeunes entreprises ayant au centre de leurs modèles et stratégies économiques la donnée qu'ils collectent, analysent et mettent au service de leurs activités.

Nous sommes dans un changement de paradigme de ce qu'est une entreprise. **Les entreprises physiques reposent sur des actifs physiques, les entreprises numériques sur des actifs numériques**. Lorsqu'une entreprise est construite autour d'actifs numériques, la conception et le développement appropriés de ces actifs numériques impacteront fortement son succès.

> *« Une Data Driven Company est une entreprise qui cherche continuellement à améliorer l'ensemble des processus de l'entreprise par l'utilisation qualitative et quantitative de données, tout le temps et sur tout ». Joseph Glorieux, OCTO.*

Comme nous l'avons vu dans le chapitre précédent, certaines transformations sont nécessaires pour qu'une entreprise puisse se dire véritablement Data Driven. L'entreprise doit tout d'abord **définir quelles métriques de succès seront mesurées**, et relier ces métriques aux ensembles de données qui contribueront à leur mesure. Cette initiative permet à l'entreprise de se préparer à aligner l'exécution tactique de chaque département à la stratégie globale de l'entreprise et à mesurer les performances par rapport aux objectifs et buts définis. L'utilisation des données et des outils analytiques doit **se propager dans toute l'entreprise**. Dans le cas contraire, les objectifs seront plus difficiles à atteindre. Il est possible pour les dirigeants de stimuler l'adoption en quantifiant puis en partageant les bénéfices financiers et l'impact sur la productivité de l'entreprise. Plus l'utilisation des outils analytiques mûrit, plus l'adoption se propage et plus la **collaboration** s'améliore. La gouvernance des données est indispensable pour permettre aux employés d'avoir accès à des **informations claires, pertinentes et consistantes**. La Data Gouvernance permet d'assurer la qualité des données et facilite le suivi des conformités aux règlementations. Beaucoup d'entreprises font l'erreur de déléguer la mise en place d'un programme de Data Gouvernance aux départements informatiques, mais ces initiatives ont plus de chances de rencontrer le succès lorsqu'elles sont **prises en charge directement par les dirigeants** de l'entreprise. Pour cause, ce sont les dirigeants qui ont accès aux **données de référence à l'échelle de l'entreprise**, aussi appelées Master Data. Bien entendu, les dirigeants doivent profiter du soutien d'experts dans le domaine des données pour pouvoir déployer une stratégie de Data Governance fiable et solide.

Voici cela résumé dans une Carte heuristique (Mind Map) :

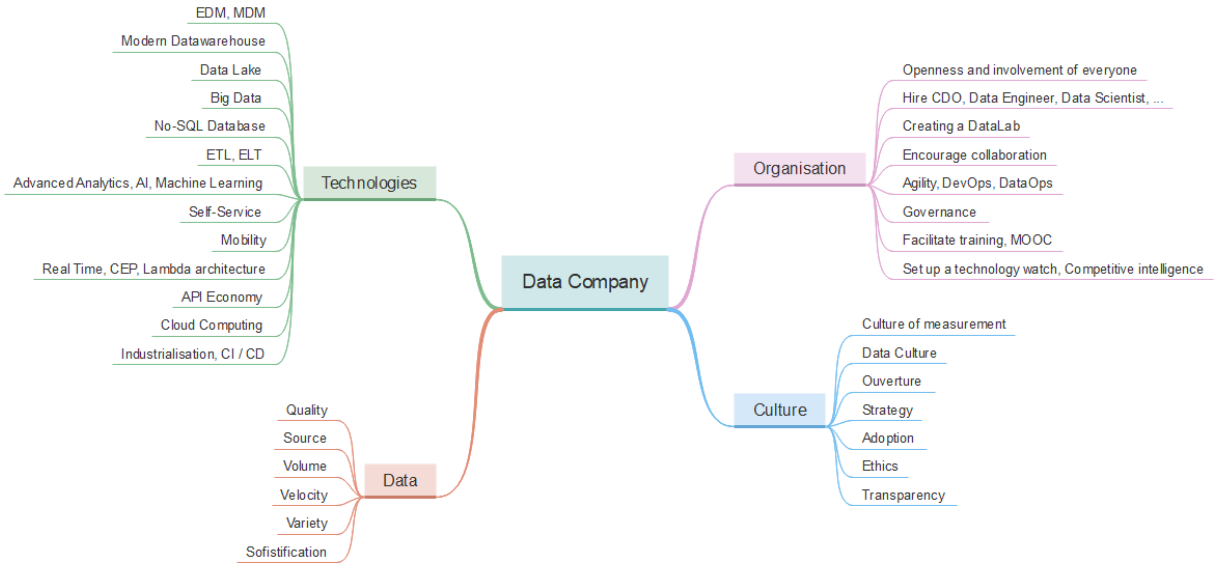
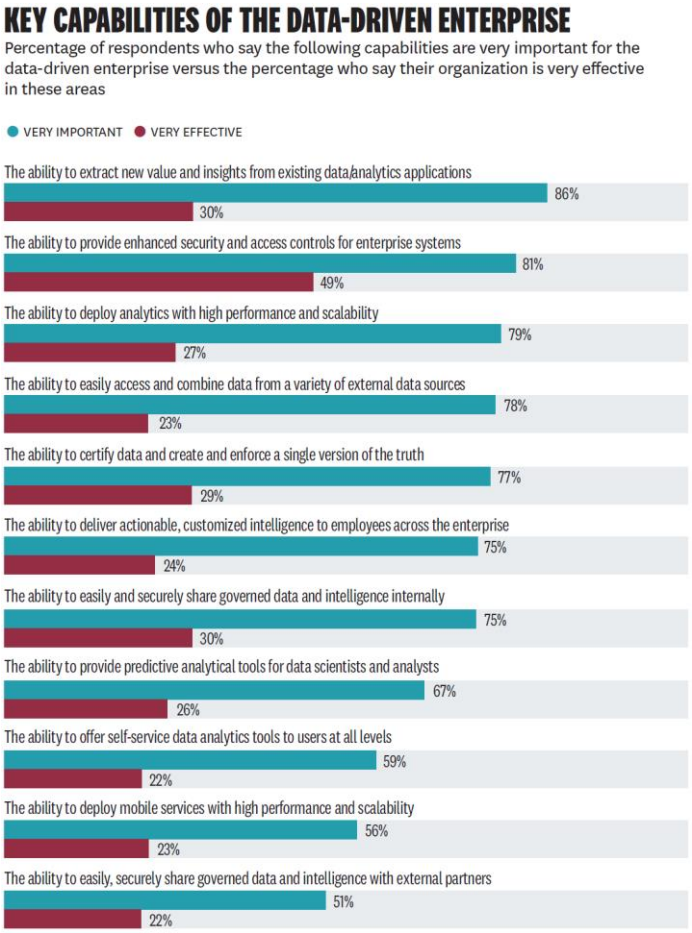
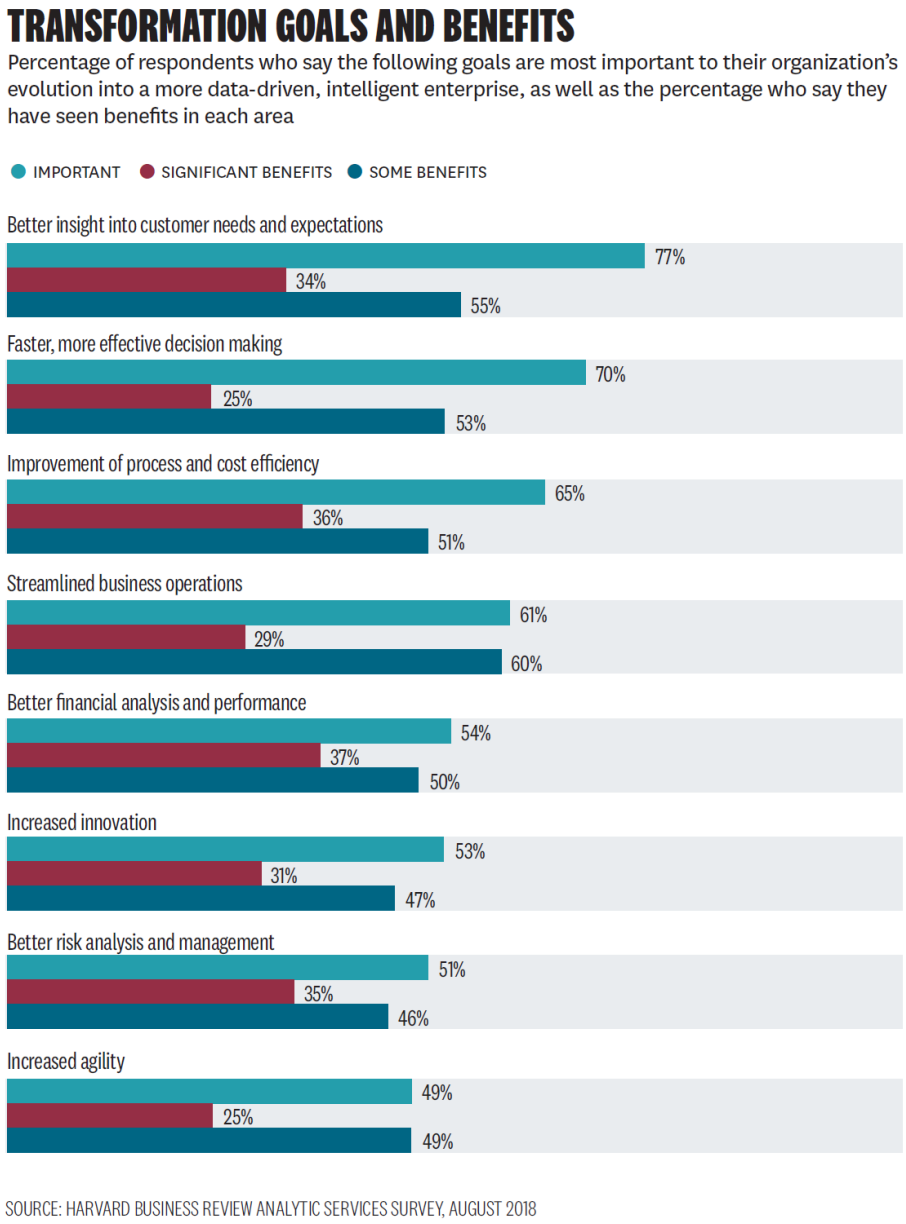


Figure 25 - Carte heuristique de la Data Company

Capacités Clefs : Une étude réalisée par Harvard Business Review Analytic Services [20] révèle que l'entreprise fondée sur les données devrait avoir la capacité d'extraire de la valeur et de mieux comprendre les données existantes, fournir des contrôles de sécurité et d'accès améliorés, déployer des outils d'analyse performants et évolutifs, accéder aux données provenant de diverses sources de données externes et les combiner, créer une version unique de la vérité et fournir des informations personnalisées et exploitables dans l'ensemble de l'organisation :



Objectifs de transformation et bénéfices : Un tiers des entreprises interrogées déclarent avoir amélioré de manière significative la compréhension de leurs clients, leurs processus et leurs coûts, tandis qu'un quart ont déclaré avoir constaté une amélioration significative de leur processus décisionnel.



Contrairement à une entreprise traditionnelle qui traite ses données comme des passifs, les Data Companies exploitent de manière agressive les données en tant qu'actifs fondamentaux. Ils génèrent des retours continus en instrumentant délibérément leurs entreprises pour collecter des données puis expérimenter pour développer de la valeur.

Alors que de plus en plus d'entreprises cherchent à devenir Data Driven, elles sont nombreuses à échouer dans la transformation de leurs données en informations exploitables. **Les principales causes de ces échecs sont souvent le manque d'organisation, de personnes qualifiées, d'outils ou de technologies nécessaires pour atteindre les objectifs fixés**.

Les étapes : Le processus pour devenir une entreprise orientée autour de la donnée n'est pas aisé, c'est un travail quotidien. Christopher S Penn a défini 5 étapes :

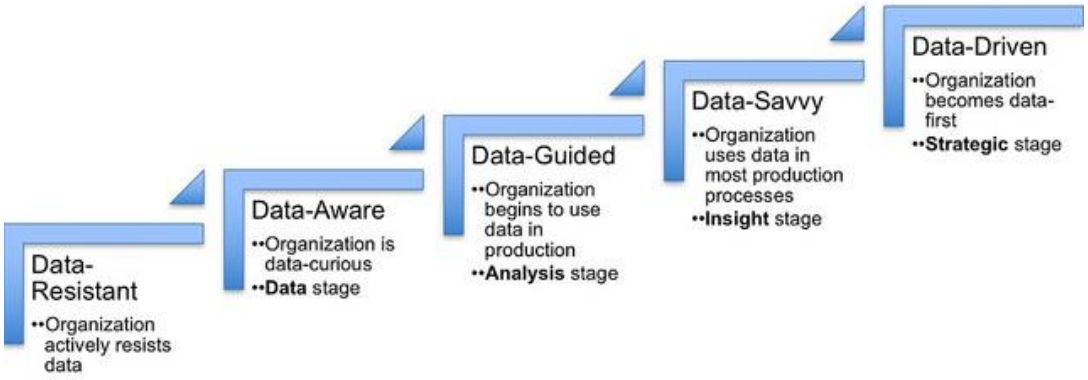
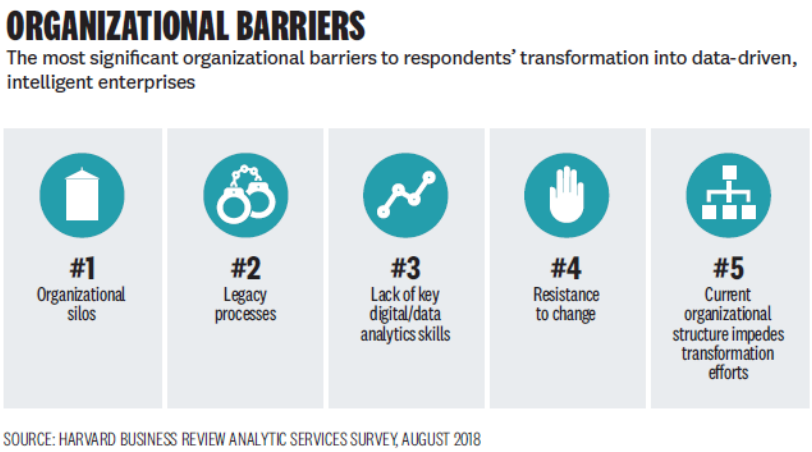


Figure 26 - The evolution of the data-driven company - Copyright Christopher S. Penn | @cspenn | cspenn.com

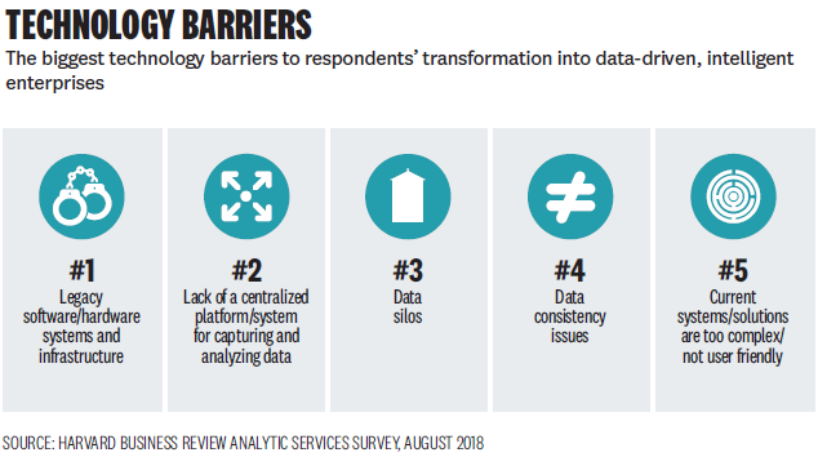
1. Data-resistant (Résistante) : La devise de l'entreprise résistante aux données est la suivante : « Nous l'avons toujours fait ainsi ». Voici quelques raisons les poussant à réagir ainsi : Les données peuvent révéler des problèmes de performances cachés, les données peuvent montrer que l'organisation a une stratégie mal alignée …
2. Data-aware (Consciente) : L'entreprise consciente de l'importance des données connait l'existence de données dans ses systèmes d'information et comprend que ces données ont une valeur implicite même si celles-ci n'ont pas encore été activées. La transition vers l'étape d'après repose essentiellement sur la volonté de débloquer la valeur des données.
3. Data-guided (Guidée) : L'entreprise guidée par les données s'efforce d'extraire la valeur des données en analysant l'historique. La société guidée par les données libère ainsi la valeur tactique de ses données : "Ne refaisons plus cela".
4. Data-savvy (Avertie) : L'entreprise qui maîtrise les données se rend compte que la valeur des données n'est pas simplement tactique, les données peuvent être un atout stratégique. Pour développer cette valeur stratégique, l'entreprise férue de données continue d'investir dans le Quoi : « Combien ai-je vendu le mois précédent ? », mais se tourne ensuite vers le Pourquoi : « Pourquoi les ventes ont-elles chuté au dernier trimestre ? », « Pourquoi les consommateurs ont-ils acheté moins de notre produit ? » et développe ainsi des connaissances.
5. Data-driven (Conduite) : La société axée sur les données associe des données, des analyses et des idées pour répondre à la question « Que faire ensuite ? ». **Grâce à l'utilisation des données à tous les niveaux, dans tous les secteurs de l'organisation, la société axée sur les données adopte les données comme ressource stratégique permettant de valider chaque décision importante.** Dans une organisation véritablement axée sur les données, chaque réunion de planification commence par des données et aucune décision n'est exécutée sans une structure de gouvernance permettant de collecter et de mesurer la décision.

On le voit, même si la dynamique doit être initiée au plus haut dans l'entreprise, sans une adhésion à tous les niveaux, une organisation ne peut pas devenir véritablement axée sur les données. Le sondage de Harvard Business Review Analytic Services déjà cité [20] révèle les barrières suivantes.

Les barrières organisationnelles : Selon les entreprises interrogées, les barrières organisationnelles les plus importantes à l'évolution des entreprises axée sur les données sont les silos organisationnels, les processus hérités, le manque de compétences en matière d'analyse et la résistance au changement.



Les barrières technologiques : Alors que certaines organisations vont dans la bonne direction au-delà de l'analyse de données manuelle, ad hoc et non normalisés, la plupart ont encore du chemin à faire pour devenir des entités pilotées par les données. Les systèmes et infrastructures hérités figurent en tête de liste des barrières technologiques pour l'entreprise, suivis de l'absence d'une plateforme centralisée, de silos de données et de problèmes de cohérence des données.



Une étude menée en 2017 par TDWI sur 264 répondants montre à peu près les mêmes barrières, on y retrouve en plus le manque de soutien des dirigeants d'entreprise, la sécurité, le manque de qualité des données ainsi que le manque d'accompagnement en formation des utilisateurs : [29]

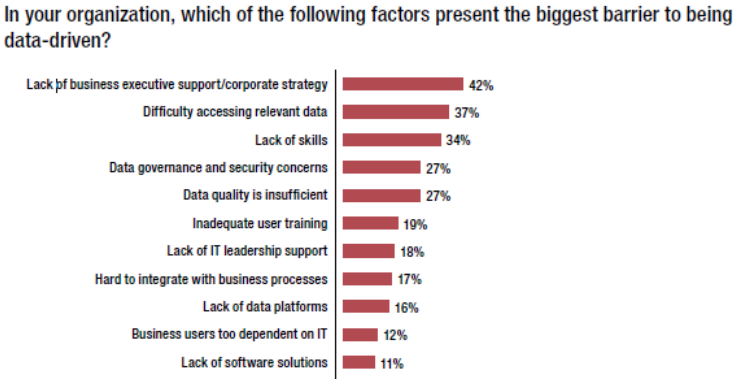


Figure 27 - Biggest barrier to being Data Driven [9]

Dans une étude Kaggle, les professionnels de l'intelligence artificielle citent les principaux défis dans leur travail : Mauvaise qualité des données, manque de talents dans l'entreprise et manque de support au niveau management et financier.

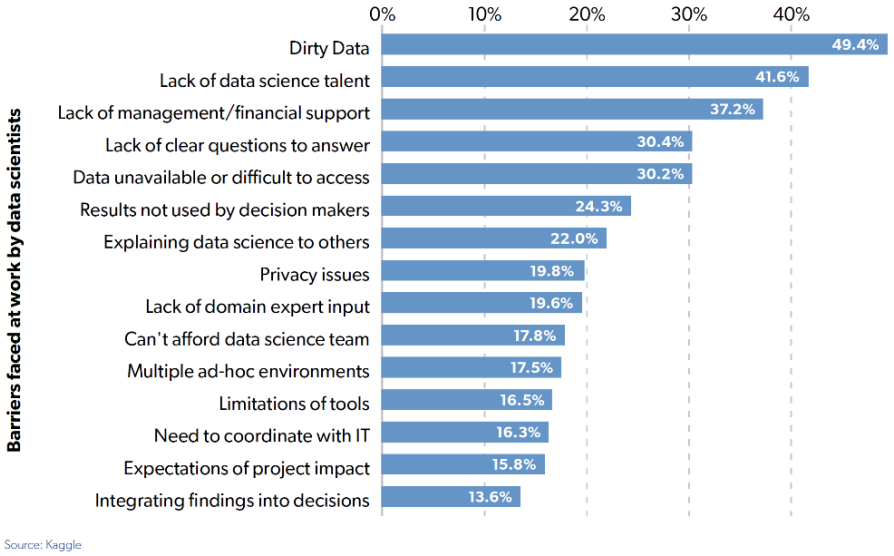


Figure 28 - Obstacles rencontrés au travail par les scientifiques des données [5]

Dans une Data Company, tout actif numérique est une source de données pour la prise de décision.

En devenant Data-Driven, c'est-à-dire en mettant l'analyse des données au cœur de sa stratégie, les entreprises se donnent entre autres le moyen de mieux comprendre les attentes de leurs clients et de capter les signaux faibles du marché. De plus, cela leur permet une gestion optimisée de leurs ressources, une meilleure collaboration entre les services, des opportunités d'adapter les produits et services en fonction des réactions du marché, la mise en place de mécanismes d'ajustements des prix agiles, la création de nouveaux services ou encore une meilleure satisfaction client. Cette promesse exige cependant que les entreprises se transforment et apprennent à travailler de façon transverse afin de mieux exploiter la donnée disponible.

Le processus visant à devenir une Data Company est multifactoriel, le rendant assez complexe à implémenter. Si plusieurs entreprises ont adopté une stratégie de valorisation de la donnée client, encore peu d'entreprises tirent parti de tous les leviers d'innovation offerts par l'exploitation de la data. En effet, cela implique un changement de culture, l'acquisition de nouvelles compétences et l'adoption d'une stratégie de gouvernance de données solides permettant d'assurer la qualité des données. Les employés doivent avoir une culture de la donnée et être en capacité de comprendre les gains liés à son utilisation. Cela demande également une prise de conscience sur les données exploitées.

Aussi et en vue des différentes barrières à la mise en place d'une entreprise pilotée par les données, il convient d'effectuer ces étapes petit à petit et de définir les métriques permettant de suivre l'avancement de ces étapes.

3.2 – Exemple de réalisations

Examinons dans ce chapitre quelques exemples de réalisations quant à la mise en place d'une stratégie Data centrique.

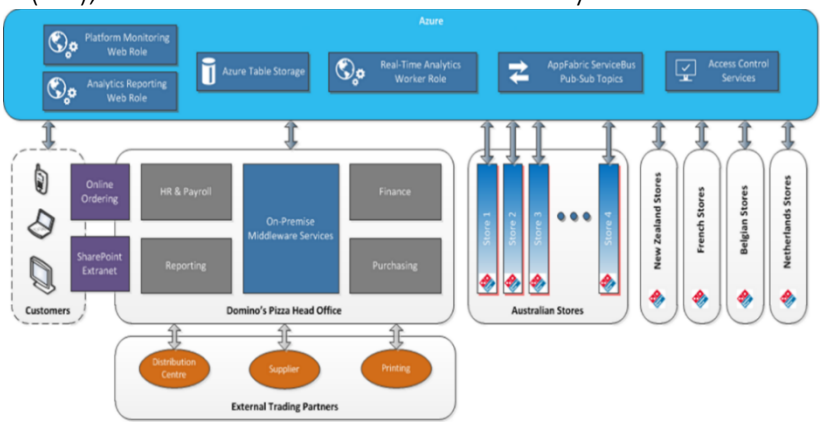
Domino's Pizza est l'une des plus grandes marques de pizza au monde. En 2010, l'entreprise perdait des parts de marché considérables du fait d'un modèle vieillissant. L'entreprise devait absolument réagir pour se redresser, innover, faire quelque chose d'audacieux pour se différencier dans un secteur déjà saturé.

Selon le magazine Forbes, la transformation numérique de Domino's nécessitait les aménagements suivants [30] :

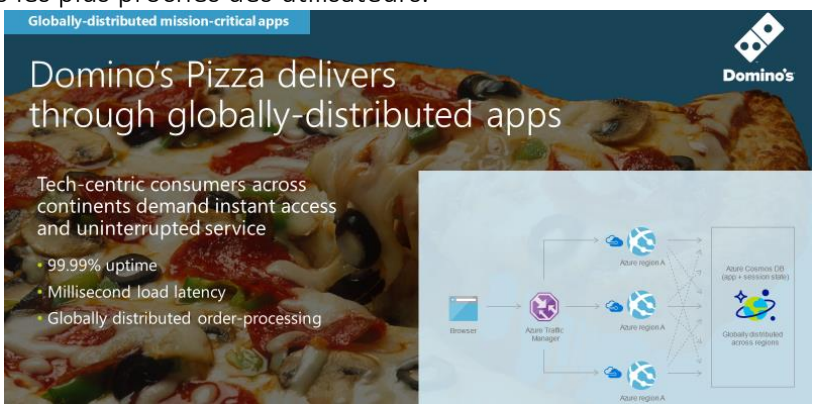
- **Changement organisationnel et mesure constante de la performance** en embauchant 400 analystes et développeurs (sur les 800 employées) afin de rendre cette transformation numérique possible.
- **Optimisation de la livraison de pizzas et utilisation des données pour atteindre efficacement de nouveaux clients, tout en gardant les clients existants.**
- Mise en œuvre d'une **culture autour de l'analyse de la donnée et vers une marque de pizza basée sur les données « Data Drive Pizza »**, ce qui a ouvert de nouvelles opportunités pour la marque dans son ensemble.

Concrètement, en plus des aspects organisationnel et culturel de l'entreprise, Domino's s'est largement appuyé sur les technologies Cloud pour réaliser sa transformation pilotée par les données. Regardons quelques-uns de ces services [31] :

- **Azure Service Bus** (Service de messagerie) : Dans le cadre de son initiative d'amélioration de ses services, Domino's a identifié le besoin de collecter, d'agréger et d'analyser les données relatives aux opérations de chaque magasin (commande, préparation et livraison de pizzas). Une plateforme de messagerie globale et peu coûteuse était nécessaire pour répondre aux exigences en matière de messagerie entre les magasins et le siège, et devait s'intégrer aux applications métiers. Domino's a choisi Azure Service Bus comme infrastructure de base pour la mise en œuvre.
- **Azure Stream Analytics** (Service d'analyse de données en temps réel, CEP) : De même, afin de traiter les communications en temps quasi réel entre les magasins et le siège social sur des informations allant des livraisons, des commandes aux performances des magasins en passant par les résultats d'enquête clients et jusqu'à l'analyse d'indicateurs clés de performance (KPI), Domino's a choisi Azure Stream Analytics.



- Cosmos DB (Base de données No-SQL) : Pour obtenir une faible latence et une haute disponibilité ainsi qu'une couverture mondiale, les données ont été déployées dans les Datacenters les plus proches des utilisateurs.



Voici les principaux facteurs opérationnels ayant incité Domino's Pizza à tirer parti d'une plateforme Cloud [32] :

- Ne pas investir dans une infrastructure informatique qu'il faudra maintenir.
- Réduire les efforts de développement et le Time to Market [33] : En privilégiant le PaaS du IaaS et en utilisant le modèle IaC [34].
- Élasticité pour soutenir la croissance et l'agilité de l'entreprise afin de supporter les pics d'activité.

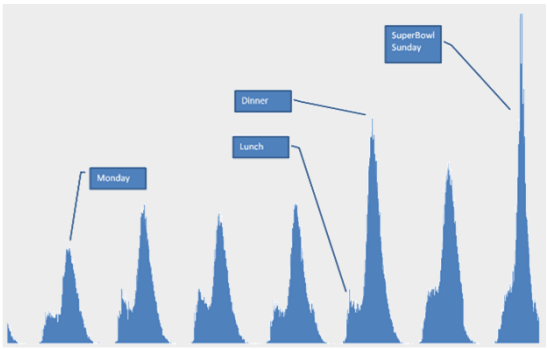


Figure 29 - Heures de pointe de commande de Pizzas aux US

Ces innovations technologiques ainsi que l'amélioration du goût des pizzas ont changé la façon dont les clients commandent. Mais la transformation numérique ne consiste pas seulement à collecter plus de données. Les entreprises doivent se transformer et pour ce faire elles doivent exploiter toutes les connaissances à partir de leurs données. **Pour réussir à long terme, une entreprise moderne doit prioriser l'infrastructure et les ressources nécessaires afin d'exploiter et extraire les données pour en tirer les meilleurs bénéfices.** Ainsi Domino's a capitalisé sur ces données pour élaborer une nouvelle stratégie autour des médias sociaux et téléphones mobiles, en connectant les clients à leur produit « Commandez vos pizzas Domino's avec un emoji ». L'entreprise a aussi découvert des informations sur les attentes de ses clients et notamment en matière de qualité : Produits frais, chauds et livraison rapide. Ces éléments ont notamment conduit à un véhicule de livraison emblématique, Adweek l'a qualifié de Batmobile de Cheese Lover's.

[33] Le Time to Market est une expression anglo-saxonne pour signifier le temps de mise sur le marché, correspondant à la durée de développement et de construction d'une offre commerciale ou d'un produit. Il est intéressant de réduire au minimum ce délai pour devancer ses concurrents et augmenter sa rentabilité. Le Time to Market est donc un enjeu stratégique majeur pour les entreprises.

[34] Infrastructure as Code (IaC) est le fait de gérer et provisionner des machines au sein d'un centre de données uniquement à l'aide de fichiers de définition plutôt qu'une configuration manuelle, à travers des interfaces interactives ou physiquement.

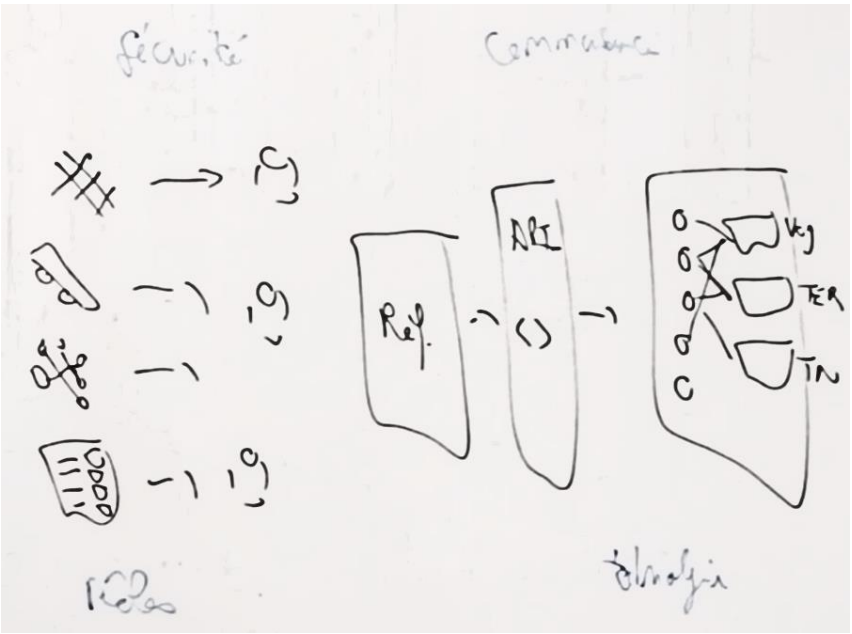
SNCF : La Société nationale des chemins de fer français est l'entreprise ferroviaire publique française. Pour la SNCF, la donnée est le levier de la performance industrielle au service des clients. À ce titre, la SNCF se dit être une « Data Driven Company » dans laquelle le patrimoine de données contient entre autres :

- Les données historiques comme les horaires, les données industrielles archivées de toutes les interventions sur les 15 000 trains, les 30 000 kilomètres de voies, dans les 3 000 gares.
- Les données que les clients confient pour améliorer l'expérience de voyage de bout en bout, à travers les services proposés en gare, à travers l'expérience à bord (connectivité 3G/4G, Wifi…), à travers une offre de mobilité multimodale…

Pour Benoît Tiers Directeur général e.SNCF, les nouvelles technologies, comme l'IoT ou le Big Data, permettent à SNCF d'innover : « *Nous avons désormais la capacité de gérer une quantité de données beaucoup plus importante, de les croiser pour proposer de nouveaux services* » [33]. Aujourd'hui, la problématique est de mettre toutes ces données au service des décisions et du pilotage de l'entreprise, de gagner en efficacité collective et en performance, avec toujours plus de sécurité et de régularité pour plus de mobilité. « *La data est assurément la clé de la mobilité de demain* » affirme M. Tiers. Ainsi, la gouvernance des données à la SNCF s'articule suivant ces 4 axes [34] :

- La sécurité et la conformité : Règlementations, hautes disponibilités, plan de reprise d'activité, …
- La connaissance et la maitrise des données : Data Stewards, Qualité des données, …
- La Technologie : Big Data, Cloud, …
- Les rôles : CDO, Data Engineer, Data Scientist, …

Dans ce schéma, les données sont générées par les infrastructures, les matériaux roulants, les transports, les utilisateurs... Ces données structurées sont sauvegardées dans des bases de données puis inscrites dans un référentiel pour garantir leur intégrité et unicité. Enfin, les APIs exposent les données aux différentes applications (TGV, TER, Transilien).



La SNCF s'est largement appuyée sur les technologies Cloud pour réaliser sa transformation pilotée par les données. En effet la SNCF entend basculer 60% de son patrimoine applicatif dans le Cloud d'ici à la fin 2020 [35]. A ce titre AWS, IBM et Azure ont été retenus :

- Avec AWS, la SNCF exploite les services Cloud de gestion de données, à commencer par Aurora et DynamoDB.
- Avec IBM, le groupe s'appuie sur Watson pour motoriser son application d'affichage d'informations dans les gares, EVA.
- Avec Azure, la SNCF a mis en place une stratégie IA et une plateforme Big Data permettant de centraliser quelque 170 To de données et fédérer 81 flux de données issues de sources différentes. Quatre Data Lakes stockent les données pour les projets de l'institution ferroviaire en respectant l'isolation des projets. La SNCF s'appuie également sur l'implémentation native de Kubernetes sur Azure pour permettre une démarche DevOps et de répondre ainsi aux enjeux cruciaux de Time to market « *La scalabilité et éviter d'avoir à construire l'infrastructure nécessaire. Cela nous permet de nous concentrer sur ce qui est important pour nous* », Raphaël Viard CTO d'e. SNCF.

Les premières expériences de la SNCF dans le Cloud ont montré des améliorations significatives en matière de livraison et d'orchestration des environnements et l'usage du Cloud a permis d'optimiser les coûts, indique aussi Raphaël Viard. « *D'emblée, nous avons aussi mis en place une pratique FinOps. Cela nous permet des économies significatives via une utilisation fine des ressources que nous mettons en place* », « *Par exemple 61 % des serveurs de la SNCF sont en arrêt programmé et ne fonctionnent qu'aux horaires de bureau. Lorsque ces ressources sont basculées dans le Cloud, cela se traduit d'emblée par une réduction des coûts de 30 %* ».

Côté organisation, la SNCF a créé des « Fabs », ce sont des centres de compétences qui accompagnent les projets digitaux SNCF. Elles regroupent des experts en méthodes ou en technologies au service de projets IoT, Big Data, Design ou Open Innovation. Au-delà des POC, les Fabs participent à l'industrialisation des innovations. Pour cela, elles misent sur la vision utilisateurs, le collectif et apportent une mutualisation des connaissances acquises lors des expérimentations et phases de prototypage. Ainsi la Fab Big Data utilise les services Azure tel que Azure Data Factory, HDInsight avec la couche de sécurisation Ranger, Databricks, …

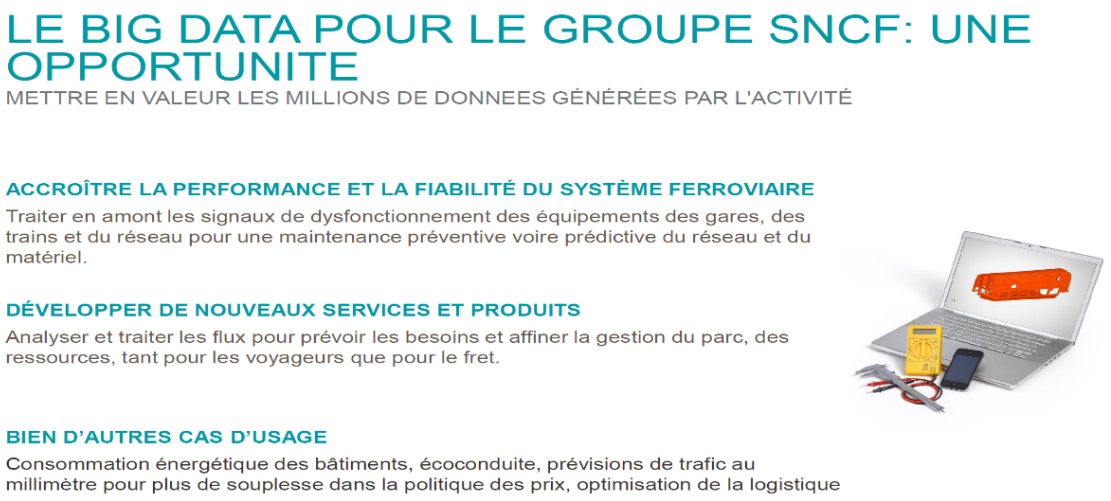


Figure 30 - Présenté par Raphaël Viard lors des MicrosoftExpériences18 [36]

Exemples de réalisation :

- Vibrato est une innovation qui consiste à récolter les données de vibration géolocalisées, via les gyroscopes et accéléromètres des smartphones embarqués dans les trains du quotidien (TER, Transilien, Intercités). Utilisée par les équipes de maintenance du réseau pour évaluer et corriger les défauts des rails, elle permet à l'entreprise d'inspecter la condition des voies ferrées et de réaliser les opérations de maintenance avant que les pannes surviennent (Maintenance Prédictive). Présentation : Projet VIBRATO.
- Data Informations Voyageurs (IV) est une application permettant de mesurer la performance d'affichage des informations sur les départs de trains dans les gares (exemple : le pourcentage de trains effectivement affichés à 20 minutes du départ). Data IV s'interface avec tous les systèmes informatiques de l'information en gare et est capable de bâtir un indicateur d'information voyageur gare par gare. Plus de deux tiers des 32 gares suivies et qui disposent quotidiennement de leur indicateur ont amélioré leurs performances d'affichage en gare. Présentation : Data IV.

Pour ces deux exemples, on voit bien la nécessité de corréler une multitude de flux de données.

Pour mettre en place sa stratégie data, SNCF s'est appuyée sur les expertises de ses collaborateurs numériques, de ses partenaires et d'un socle technologique renouvelé. Pour que l'exploitation des données soit possible, il faut les héberger de manière sécurisée, souple et robuste.

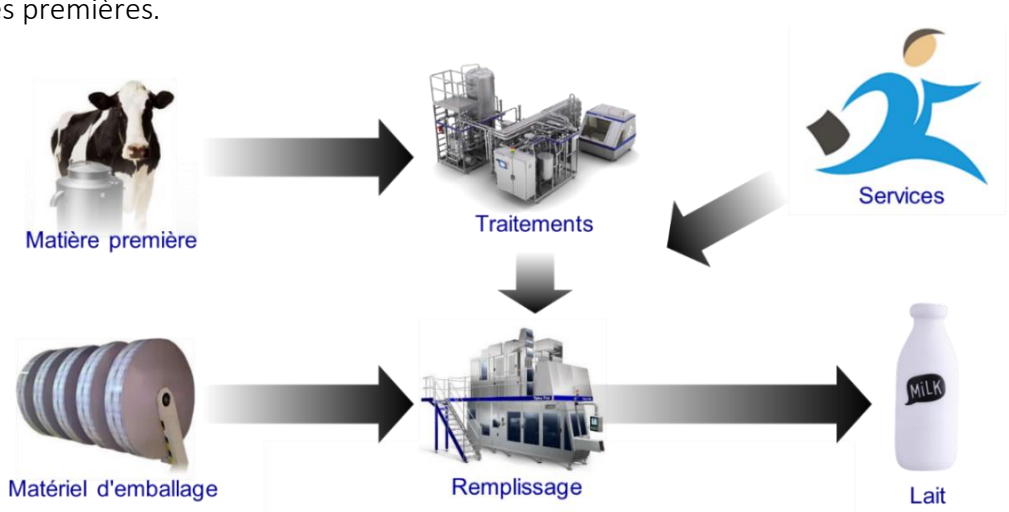
Figure 31 - #DIGITALSNCF Où en est la transformation numérique de SNCF ? [37]

SNCF a fait le choix d'héberger la moitié de ses applications sur le Cloud pour lui permettre de répondre, avec une très grande flexibilité et fiabilité, aux évolutions de la demande, comme dans le cas de l'application SNCF. Dans le cadre de la stratégie data, les applications du groupe SNCF s'échangent des données de manière simple et robuste. L'enjeu est aussi de pouvoir mettre à disposition certaines de ses données à l'externe via des API publiques. A l'image l'API SNCF qui fournit, depuis juin 2015, une large gamme de services autour des horaires des trains. Aujourd'hui, les différentes API du groupe sont centralisées dans une plateforme, ce qui accélère les développements des applications métiers.

3.3 – Étude de cas

Avant de présenter l'étude de cas, commençons par brièvement présenter l'entreprise fictive Contoso qui nous servira de support. Nous avons choisi d'utiliser une entreprise fictive pour plusieurs raisons : D'une part la confidentialité des informations et d'autre part, cela nous permet de présenter plusieurs aspects constatés sur différents projets et clients.

L'entreprise Contoso est un fabricant mondial d'équipements de conditionnement et de transformation des aliments. Contoso propose des machines de conditionnement, de remplissage et de traitement de produits alimentaires liquides telles que produits laitiers et jus de fruits, fromages et crèmes glacées. Contoso fournit à ses clients des systèmes de traitement, d'emballage et de distribution conçus pour optimiser l'utilisation des ressources et des matières premières.



Contoso souhaite transformer le secteur de la fabrication de produits alimentaires en intégrant des services numériques. L'entreprise souhaite ainsi transformer son modèle commercial de fournisseur d'équipement en société de services technologiques offrant à ses clients un aperçu en temps réel du fonctionnement de ses machines. Contoso entend accroître sa part de marché mondial et faire économiser de l'argent à ses clients grâce à une utilisation plus efficace de l'information. Cela inclut des analyses avancées, permettant aux employés sur le terrain de disposer d'informations sur l'état de dégradation des machines avant leur détérioration complète, une maintenance conditionnelle basée sur le franchissement de seuils prédéfinis. En tant que partenaire dans cette étude de cas, nous allons aider Contoso dans sa transformation numérique autour d'une architecture data centrique.

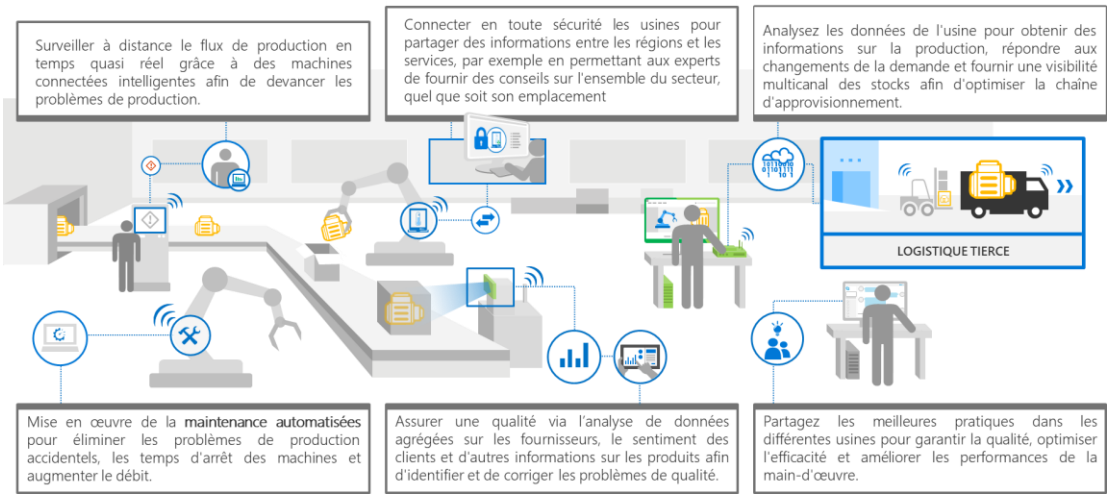
En analysant l'entreprise et ses problématiques, nous avons identifié les axes de transformation numérique suivants :

- **Clients** : L'entreprise doit offrir à ses clients des expériences personnalisées, riches et connectées. Exemple : Permettre de suivre en temps réel une chaine de remplissage.
- **Employés** : L'entreprise doit permettre à ses employés de suivre le rythme de ses clients, en collaborant efficacement pour répondre à leurs besoins avec agilité. Exemple : Fournir à ses employées les outils adéquats permettant de remédier aux pannes matérielles le plus rapidement possible et dans les meilleures conditions.
- **Opération** : L'entreprise doit pouvoir augmenter le flux d'informations dans l'ensemble de ses opérations commerciales, synchroniser ses processus métier et améliorez ses interactions avec les partenaires et sa chaîne logistique. Exemple : Anticiper les tendances en examinant les données historiques sur la production et la distribution.
- **Innovation** : L'entreprise doit pouvoir recueillir des informations sur l'utilisation de ses produits, concevoir des fonctionnalités innovantes et collaborer avec une équipe de développement pour améliorer les produits et en développer de nouveaux.

Les sujets autour de la transformation numérique de l'entreprise Contoso sont nombreux. **Dans un souci de chaine de valeur, pensons à la solution business avant de considérer la plateforme technique** aussi, focalisons-nous sur le suivi informatisé du conditionnement et la maintenance prédictive. Les objectifs du projet sont nombreux :

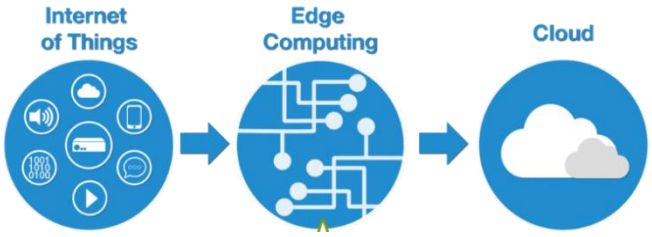
- Permettre une analyse descriptive des évènements d'une usine « Que s'est-il passé hier ? », « Combien de bouteilles ai-je emballées ? ».
- Permettre une analyse en temps réel : « Qu'est-il en train de se passer ? », « Où en est-on par rapport à hier ? ».
- Permettre une analyse diagnostique « Pourquoi c'est arrivé ? ».
- Permettre une analyse prédictive « Que va-t-il se passer ? », « Au regard des données historiques, combien de bouteilles vais-je pouvoir emballer aujourd'hui ? ».
- Permettre une analyse prescriptive « Que puis faire pour éviter que la panne se reproduise ? ».

Le but est d'augmenter l'efficacité des processus industriels en limitant le nombre d'interventions humaines tant au niveau de la maintenance que sur la gestion des pannes. En anticipant des évènements négatifs, l'analyse prédictive va permettre d'élaborer des stratégies résilientes, anticiper les pannes et par la même occasion réaliser des économies.



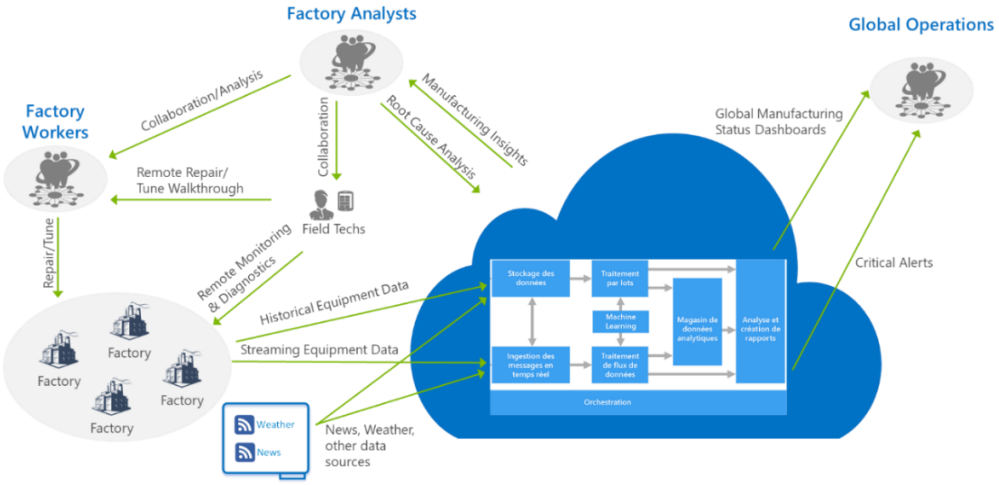
Les machines de remplissage Contoso sont installées sur les sites de clients du monde entier (plus de 5000 machines). Les périphériques connectés sur ses machines sont composés de capteurs qui détectent, mesurent la lumière, la chaleur, le mouvement, l'humidité, la pression, … et les convertissent sous une autre forme d'impulsions électriques. Le projet doit englober les données de diverses sources internes issues de capteurs d'environnement dans l'usine, de capteurs sur les machines de traitements, de remplissage et d'emballage. Mais aussi des sources externes comme l'arrivage des matières premières, les conditions météorologiques … Le projet appuiera la vision future de Contoso consistant à élaborer des solutions autour des données.

Dans un souci de scalabilité, de reproduction de l'architecture du projet à l'international (Infrastructure as code), de Time To Market, … le Cloud s'impose comme une évidence, tout comme l'utilisation de services plutôt que de plateformes (Serverless). Il conviendra de comparer les différents Cloud Providers en prenant en compte le coût des services, leurs SLAs[35], leurs certifications sécuritaires, la localisation des Datacenters, leurs couvertures mondiales, etc., les connaissances des développeurs d'une équipe sont à prendre en compte afin de maximiser le temps pour être opérationnel.



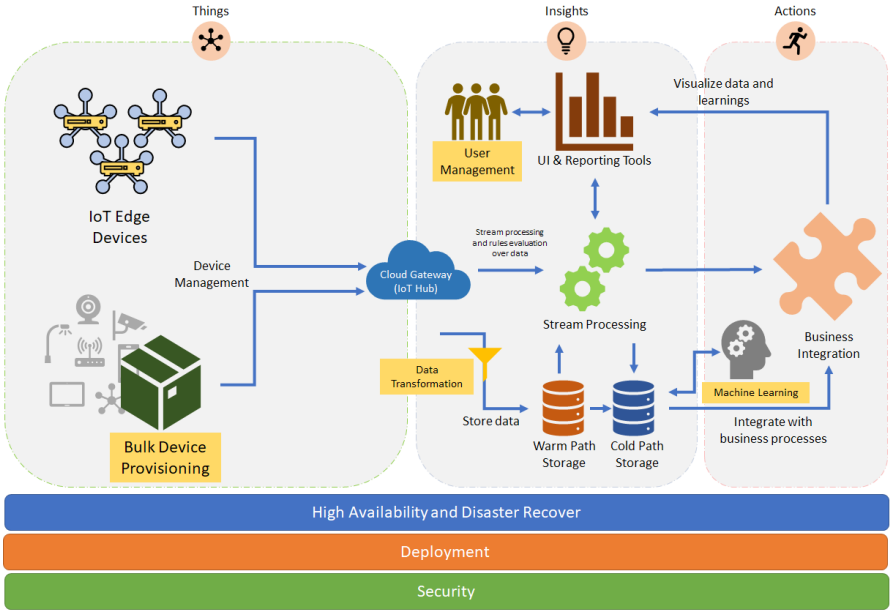
Afin de limiter des données transmises dans le Cloud et dans le but de réduire les coûts, l'utilisation du Edge Computing[36] permet de rendre plus intelligents les périphériques et réduit les données de télémétrie brutes avant leur transfert vers le Cloud. Seules les informations pertinentes, agrégées, filtrées et sécurisées seront transmises. En déplaçant une partie de la charge de travail aux périphériques connectés cela permet aussi aux équipes sur site de réagir plus rapidement aux changements de statut. Différentes passerelles IoT servent de point de connexion entre le Cloud, les contrôleurs, les capteurs et les périphériques intelligents dans l'usine.

Storyboard de la solution :



[35] Le Service Level Agreement, ou SLA est un contrat par lequel un prestataire informatique s'engage à fournir un ensemble de services à un client, il définit les objectifs précis et le niveau de service.
[36] L'edge Computing ou l'informatique en périphérie de réseau est une méthode d'optimisation employée dans le Cloud Computing qui consiste à traiter les données à la périphérie du réseau, près de la source des données.

Architecture conceptuelle :



Les technologies du projet :

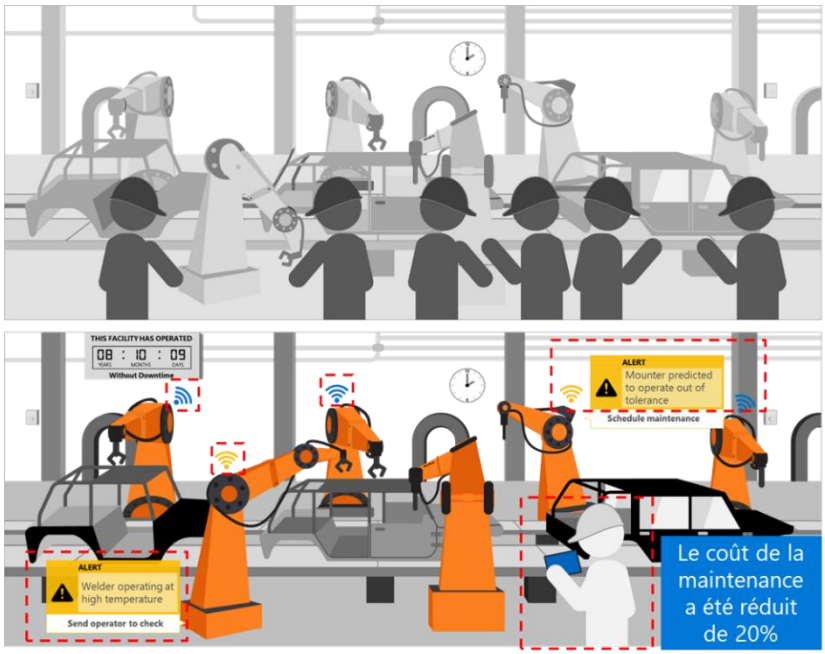
- Des objets connectés intelligents (IoT Edge) ainsi qu'un service Cloud pour les connecter, les surveiller et les gérer de manière sécurisée (Gateway, Hub, Passerelle).
- Un espace de stockage évolutif de type Data Lake.
- Un service de traitement de données d'évènements (CEP).
- Un service de Machine Learning exposant des APIs.
- Un Datawarehouse en tant que service facilitant l'analyse.
- Un service de calculs permettant le traitement en mode Batch d'un grand volume de données.
- Un service de restitution interactif disponible sur tout type de périphériques permettant la génération d'alertes.
- Un orchestrateur de la donnée permettant l'automatisation de traitements.

Les données et les évènements enrichis, filtrer, sont transmis aux passerelles via le protocole radio Z-Wave augmentant ainsi la résilience et limitant la consommation électrique des objets connectés. Les données sont ensuite transmises avec le protocole sécurisé REST/HTTPS en utilisant le format JSON car facilement lisible et le format Apache AVRO car compressé et performant [38] à un service Cloud d'ingestion de données déployé dans un Datacenter proche de l'usine. Ces données sont enrichies en temps réel via un service de CEP appelant des API de Machine Learning afin de prédire la durée de vie restante des pièces. Les données sont persistées dans un Data Lake dans leurs formats d'origine, Les résultats sont analysés en temps réel et sont stockés dans une base de données et modélisées dans un service d'analyse destiné aux utilisateurs.

Concernant la sécurité, les périphériques sont identifiés, authentifiés et les connections sont sécurisées de bout en bout (Objet → Passerelle → Cloud). Certaines données sensibles sont encryptées sur disque « Encryption at rest » ou pendant leur transfert « Encryption in motion ». Enfin, l'accès aux données est géré par une authentification Multi-factor a un Active Directory. Des systèmes de surveillance fournissent des informations sur la santé, la sécurité, la stabilité et les performances de la solution.

Concernant la haute disponibilité et la gestion de sinistre (HA/DR), la duplication de services Cloud dans différentes régions permet d'assurer et d'accroitre la disponibilité des services. Concernant la conduite et l'organisation du projet, nous avons réalisé le POC en utilisant la méthode agile SCRUM[37] afin de délivrer au plus vite une solution exploitable (MVP[38]).

Avant, l'usine subissait des interruptions de maintenance non prévues, des perturbations dans la chaine d'approvisionnement ainsi qu'une utilisation non optimale des ressources de fabrication. Désormais, l'usine est connectée et utilise des algorithmes permettant de prévoir la maintenance et l'optimisation des lignes de production en fonction de la demande.



En proposant à ses clients de monitorer la chaine de conditionnement via des objets intelligents afin de prédire la maintenance des équipements, Contoso permet d'accroître la valeur ajoutée de ses services. Nous avons vu l'importance de la culture de la donnée dans l'entreprise. A ce titre Contoso offre un service de formations autour de l'analyse des données à ses clients et collaborateurs s'assurant ainsi qu'ils disposent du savoir-faire nécessaire pour activer et analyser les données générées et ainsi prendre les meilleures décisions dans les plus courts délais.

Avec les technologies, la culture et l'organisation Data centrique, Contoso a tous les atouts pour tirer le meilleur avantage de l'Industrie 4.0. L'Industrie 4.0 fait référence à la tendance actuelle en matière d'automation et d'échanges de données relatives aux méthodes et technologies industrielles. Cela comprend les systèmes cyber-physiques, l'Internet des Objets, le Cloud et Cognitive Computing, le Digital Twin. [39]

En devenant de plus en plus pilotées par les données et en mettant davantage l'analyse des données au cœur de leur stratégie, les entreprises comme Contoso peuvent associer l'automation aux systèmes informatiques. Ses systèmes d'information sont constamment enrichis de données disponibles en direct permettant de conduire les opérations de manière toujours plus efficace et productive.

[37] Le Scrum est un Framework lié aux méthodes agiles de gestion de projet.
[38] Le produit minimum viable (ou MVP de l'anglais Minimum Viable Product) est une stratégie de développement de produit dont l'objectif est de valider l'adéquation entre la proposition de valeur et le marché.

Conclusion

Que signifie être piloté par les données ?

La solution n'est pas de disposer uniquement de technologies récentes ni d'avoir une équipe de Data Scientist chevronnés. Comme exposé dans le mémoire il faut aussi une organisation et une culture autour de la donnée « **Data Company = Technologie + Organisation + Culture** ».

Nous l'avons vu, les données affectent toutes les parties de l'organisation, les ressources numériques et les produits de données sont de plus en plus répandus.

Affectent-ils le mode de fonctionnement des entreprises ?

Nous pensons que c'est le cas et que les entreprises les plus avancées modifient de plus en plus la manière dont elles gèrent les processus critiques de leur activité en tirant parti des données.

Le fait que les entreprises doivent s'organiser de plus en plus autour des données génère plusieurs défis organisationnels, culturels et technologiques.

Les progrès de l'intelligence artificielle et la quantité d'informations générées quotidiennement dans le monde ouvrent de nouvelles possibilités qui mènent à de nouvelles connaissances. De même, la puissance de calcul des machines combinées à la créativité de l'homme permet de repousser les limites de la connaissance actuelle. Un projet ou une initiative de transformation numérique doit s'appuyer sur tout ou partie des technologies **SMACS** (Social, Mobile, Analytics, Cloud, Security), d'intelligence artificielle, de réalité virtuelle, d'IoT ou de Big Data par exemple.

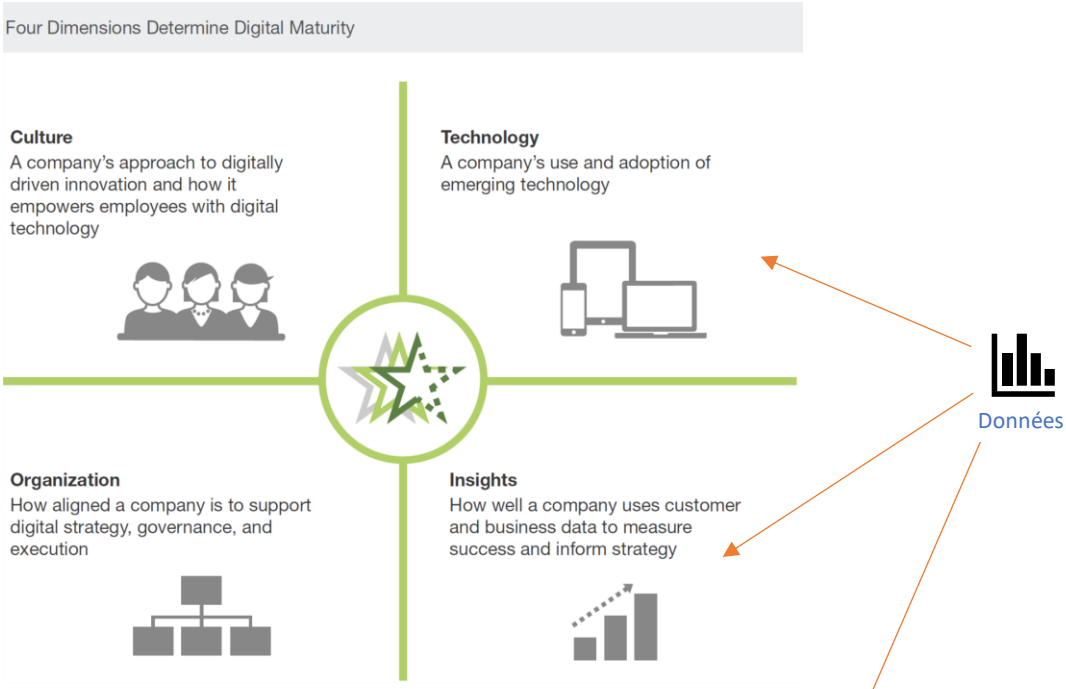
Alors que l'entreprise classique construit son système d'information en se basant sur son Business Model[39] et ses processus existants, l'entreprise digitale construit ses processus et son Business Model à partir de son système d'information lui permettant ainsi d'observer et d'analyser les comportements de ses clients. De même, une entreprise classique fait encore de la Business Intelligence en analysant les données du passé. Une entreprise digitale fait du « Operational Intelligence » en analysant de manière dynamique et en temps réel les données de ses clients pour réagir très rapidement à la situation avant que l'opportunité (Business Moment) soit passée.

La transformation digitale s'appuie aujourd'hui sur des technologies plus ou moins matures. Les enjeux tournent autour de la transformation de l'expérience client et de la transformation des outils de production. L'IT ne doit pas être seulement un support au processus métiers mais il doit fusionner avec celui-ci (DevOps, DataOps). **Ainsi la DSI n'a pas seulement un rôle technique, elle a un rôle stratégique dans l'entreprise digitale**. Elle doit proposer de nouveaux business modèles et être source d'innovation pour l'entreprise. Pour lui faciliter la tâche, le Cloud, les applications en mode SaaS, les API la libère de précédentes contraintes comme le déploiement et la maintenance d'infrastructures physiques, …

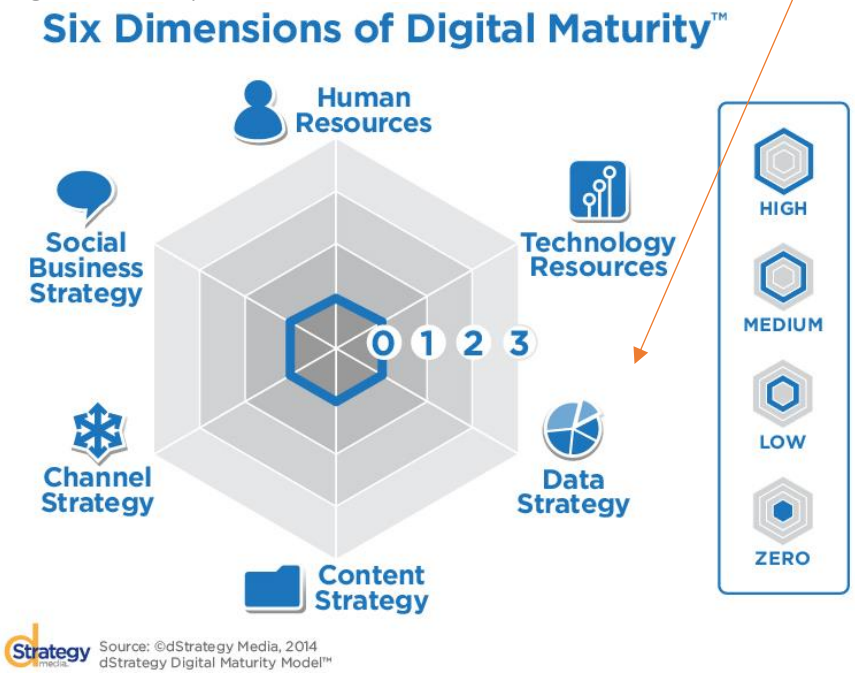
[39] Le business model a pour fonction de décrire la manière dont une entreprise crée de la valeur et assure ainsi sa propre pérennité.

Avant de répondre à la question de recherche posée au début du mémoire « La donnée est-elle au cœur de la transformation numérique des sociétés ? », regardons les différents systèmes de maturité numérique élaborés par différentes sociétés de conseils permettant de qualifier l'état de transformation numérique des entreprises.

- Forrester Digital Maturity Model 5.0 :



- dStrategy Digital Maturity Model :



Un point commun entre tous ces systèmes de calcul de maturité de la transformation numérique : La prépondérance de la donnée, qui aurait donc un rôle primordial dans la transformation numérique des entreprises.

Cependant, même si la culture axée sur les données est un facteur d'efficacité qui favorise la création de valeur et soutient la croissance des entreprises, Alan Duncan et Frank Buytendijk du cabinet Gartner expriment dans un rapport certaines limites au tout axé sur ou dirigé par les données [40] que nous avons étayées :

- Lorsqu'il n'y a pas assez de données disponibles, par exemple lors de l'entrée d'un produit sur un nouveau marché. Dans ces cas, il est nécessaire de recourir à des techniques alternatives telles que la planification de scénarii ou la modélisation. L'absence de données ne signifie pas que l'assertion est incorrecte, l'absence de données justificatives ne doit pas toujours être interprétée comme un point de preuve négatif. Rappelons qu'il est souvent préférable d'avoir peu de données de qualité que beaucoup de données de mauvaise qualité.
- Une décision basée sur des données ne permet pas de discerner quelque chose de juste ou de non juste d'un point de vue moral ou éthique. L'utilisation des données n'améliore pas nécessairement toutes les décisions de l'entreprise.
- Un focus excessif sur les métriques ou l'obsession de la mesure si elle est poussée à l'extrême peut provoquer la "paralysie de l'analyse" empêchant ainsi l'innovation. Il est nécessaire d'assumer un certain niveau de risque, d'être flexible, car si trop de temps et de ressources sont investis dans l'analyse alors qu'aucune décision n'est prise, l'entreprise stagne. En effet, il peut arriver que la solution claire à 100% s'impose par les données dans un temps imparti. Toutes les décisions ne peuvent pas être étayées par des données.
- Une étude a montré que neuf dirigeants sur dix trouvent le moyen d'ignorer les données si elles contredisent leur point de vue [41]. Comme quoi, même si les données sont présentes, sans culture de la donnée au sein de l'entreprise (curiosité, remise en cause des acquis, …) elle est vaine.
- Il peut exister un effet d'amplification : Si les employés d'une entreprise sont focalisés sur leurs objectifs personnels, leur bonus par exemple, alors ils ne verront que cela au détriment d'autres objectifs plus larges mais qui ont potentiellement plus de valeur pour l'entreprise. Cet effet peut être appelé le « Cobra Effect » [40].

Pour être une entreprise basée sur les données, il ne suffit pas d'avoir la technologie ni l'organisation adéquate, il faut avant tout accepter le besoin d'un changement culturel, savoir passer outre les limitations et savoir accompagner le changement dans l'entreprise. Ce n'est que de cette manière que les barrières pourront être franchies et que le potentiel sera effectif.

[40] L'effet cobra se produit lorsqu'une tentative de résolution d'un problème aggrave le problème. Il trouve son origine dans une anecdote créée à l'époque de la domination britannique de l'Inde coloniale. Le gouvernement britannique était préoccupé par le nombre de serpents cobra venimeux en Inde, à Delhi. Le gouvernement a donc offert une prime pour chaque cobra mort. Au début, cette stratégie était une réussite, car un grand nombre de serpents ont été tués pour la récompense. Cependant, des personnes ont commencé à élever des cobras pour gagner de l'argent. Lorsque le gouvernement en a eu connaissance, le programme de récompenses a été supprimé, ce qui a incité les éleveurs de cobra à libérer les serpents désormais sans valeur. En conséquence, la population de cobra sauvage a augmenté.

Malheureusement, être piloté sur les données ne garantit en rien le succès. En effet, les stratégies les plus réussies peuvent être copiées par des concurrents et alors que les C-Level dirigent l'entreprise, si la définition de la vision et de la stratégie n'est pas bonne même si elle est fondée sur les données, elle peut mener l'entreprise a la faillite. Exemple de l'entreprise Tesco, un groupe de distribution international, qui a souvent été salué comme un modèle de gestion de données [42]. Cependant sa valeur marchande s'est effondrée en 2014 du fait d'échecs opérationnels, organisationnels et culturels [43].

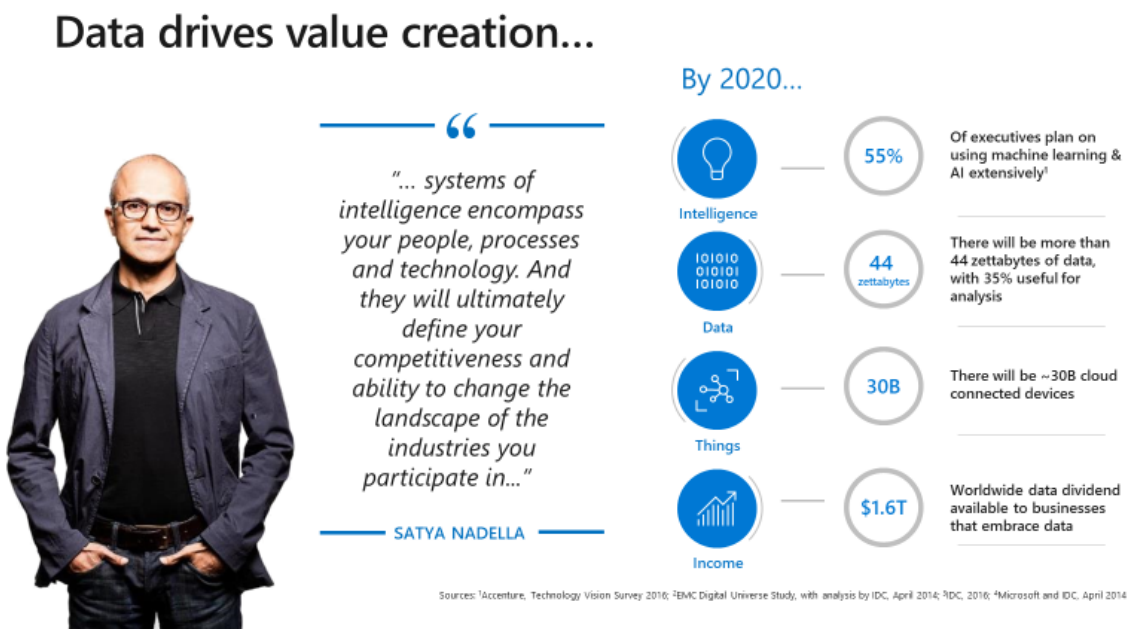
Malgré quelques contre-exemples un certain nombre d'études prouvent que le fait de s'appuyer sur des données est rentable, les entreprises prennent de meilleures décisions, innovent davantage et plus rapidement. Les entreprises où la culture de la mesure est présente, testent et savent si oui ou non leur effort fonctionne à partir d'indicateurs clés de performance. Elles sont alors plus inclusives et tout le monde peut contribuer et voir comment leurs contributions participent au succès de l'entreprise.

Nous vivons incontestablement une « révolution des données ». Nous avons vu qu'avec la dématérialisation d'un nombre croissant de processus et l'apparition de produits et services numériques totalement nouveaux, la quantité de données numériques créées augmente de manière exponentielle. Alors que ces données prennent de plus en plus d'importance pour les entreprises, les consommateurs, les États et les citoyens, leur traitement devient bon marché et les outils utilisés pour les traiter et les analyser font l'objet d'une innovation constante. En parallèle, de nouvelles politiques d'utilisation des données voient le jour et la demande de ressources humaines possédant des compétences dans le traitement des données connaît une hausse importante. La donnée est désormais le fer de lance de la révolution numérique. Comme les autres révolutions industrielles avant elle, la révolution des données transforme notre économie, notre société et nos politiques.

> *"I have an easy but firm conviction, from now on the way companies capture, manage and use information will determine profit and loss of companies" Bill Gates [40]*

La transformation numérique a mis en exergue la nécessité de développer des actifs numériques pour se connecter aux consommateurs. L'impératif d'une stratégie de données est une conséquence directe de la stratégie numérique. Les clients et employés exigent de plus en plus de produits de données afin de prendre de meilleures décisions et de bénéficier d'une meilleure expérience utilisateur. Ainsi la transformation numérique ou digitalisation consiste à mettre en place une plateforme informatique, un système d'information entre les clients et l'entreprise. Cette plateforme, dans laquelle le client est au centre, permet dans un premier temps de gérer la relation avec le client et dans un second temps de collecter des données pour mieux comprendre ses besoins et ses comportements afin de lui proposer l'offre qui lui correspond le plus. Comme le client exige tout, tout de suite, l'informatisation des processus permet de fluidifier la production de service et de biens pour répondre rapidement aux besoins.

Les données sont le carburant de la transformation et l'analyse, le moteur de l'exécution de nouvelles stratégies.



Nous sommes convaincus que les entreprises qui ne croient pas que leurs données constituent un atout et qu'elles doivent être gérées en conséquence, auront de grandes difficultés dans les prochaines années. C'est un impératif existentiel, les sociétés qui ne deviennent pas des entreprises axées sur les données seront remplacées par celles qui le seront. La donnée est bel et bien au cœur de la transformation numérique des sociétés.

Le terme Data Company n'existe pas dans le dictionnaire et il n'existera peut-être jamais car même si c'est une caractéristique indéniable des entreprises, toutes celles qui seront encore là demain seront forcément pilotées par les données. Les entreprises seront donc toutes à terme des Data Companies et le terme ne sera alors plus un différenciateur mais un point commun. Il sera ainsi bien moins utilisé qu'il ne l'est aujourd'hui. Notre sentiment est que la dimension culturelle des données et de leur analyse avancée en est à ses débuts : Nous serons de plus en plus influencés par l'intelligence artificielle et le Big Data.

Remerciements

Je remercie tout d'abord Microsoft et ma manager Véronique Mery de m'avoir offert l'opportunité de suivre la formation proposée par CentraleSupélec : Skill Evolution Program Cursus Advanced.

Merci à Microsoft France et Centrale Supélec à l'origine de ce cursus et plus particulièrement Pierre-Frédéric Rouberties, Bernard Ourghanlian, Philippe Beraud et Vincent Martin.

Merci à l'ensemble des intervenants de CentraleSupélec qui ont partagé leurs connaissances et expériences ainsi que pour la qualité de leurs interventions.

Merci à Olivier Breton, Régis Baccaro, Sophie Bismuth, Mathias Ekizian de Microsoft France pour leur partage et retour d'expérience.

Merci à la famille et plus particulièrement à Guy Casteres et Bernard Papin pour leur relecture et leurs retours constructifs.

Enfin, je tiens à exprimer ma gratitude envers ma compagne Myriam Papin ainsi qu'à mon jeune garçon de 4 mois Léo pour m'avoir soutenu et encouragé tout au long de cette formation et de s'être montrés compréhensifs durant les samedis de formation.

Bibliographie

1. Article de Armelle Gilliard dans le magazine Madmagz : EDUCAVOX
2. Livre de Lucas D. Introna “Management, Information and Power: A narrative of the involved manager”
3. Etude Infolab : Mettre la culture des données au cœur de l’organisation
4. Article de Paul Petrone sur LinkedIn : The Skills Companies Need Most in 2018
5. Edude de MMC Ventures : The State of AI: Divergence, 2019
6. Article de Susan Moore sur Gartner : How to Create a Business Case for Data Quality Improvement
7. Présentation de Bill Chamberlin : The Data Economy: 2016 Horizon watch Trend Brief
8. European Commission : Communication on data-driven economy
9. Définition de la « Data Monetization » : https://www.gartner.com/it-glossary/data-monetization
10. Interview de Clay Christensen par Harvard Business School : Disruptive Innovation Explained
11. Livre de José Antonio Martínez Aguilar “The Data Advantage - The Rise of Smart Companies: How to Create a Competitive Advantage with Data and Artificial Intelligence”.
12. Guide du Gartner : 2017 Planning Guide for Data and Analytics
13. Livre de Eric Ries (2011) “The Lean Startup”, Diagramme Build-Measure-Learn Feedback Loop
14. OECD (2013), “Exploring Data-Driven Innovation as a New Source of Growth: Mapping the Policy Issues Raised by "Big Data"”, OECD Digital Economy Papers, No. 222, OECD Publishing, Paris. http://dx.doi.org/10.1787/5k47zw3fcp43-en
15. Article de Mark Beyer, Gartner : Logical Data Warehouse
16. Article de Gregory Piatetsky, KDnuggets : Python eats away at R
17. Etude BARC : BI Trend Monitor 2019
18. Article de Christophe Parageaud : Big Data : Panorama des solutions 2016
19. Article Edureka : Learn how to Setup a CI CD Pipeline from Scratch
20. Etude de Harvard Business Review : Inflection Point Data Driven Enterprise
21. Etude de datarecrutement.fr : Etude Salaire 2019
22. Article de Cyrielle Chauwin : Le Data Lab : clé de voûte de vos projets data
23. Article Doing Data Science
24. Article de Serge Courrier : DSI et veille stratégique
25. Etude Institute for the Future et Dell : Ermerging technologies impact on society & work in 2030
26. The Official Microsoft Blog, April 15, 2014 : A data culture for everyone
27. Article neilpatel.com : How To Master A/B Split Testing Quickly
28. Hauser, John & Katz, Gerry. (1998). Metrics: You are what you measure!. European Management Journal. 16. 517-528. 10.1016/S0263-2373(98)00029-2.
29. Etude TDWI : What It Takes to Be Data-Driven
30. Article Forbes : How Domino's Transformed Into An E-commerce Powerhouse Whose Product Is Pizza
31. Brighttalk : How Azure PaaS Supports the Expansion of Domino's
32. ARCast.TV : Selling Pizza in the Cloud Domino's and Windows Azure
33. Article : La donnée, nouvelle étape de la transformation de SNCF
34. Vidéo SNCF : La gouvernance des données : Le calcul d'indicateurs

35. Article Lemagit : La SNCF va basculer 60% de ses applications dans les Clouds d'Amazon AWS, Microsoft et IBM
36. Keynote : Microsoft Experience 18
37. DigitalSNCF : Où en est la transformation numérique de SNCF ?
38. Article de Stitchdata : COPY performance of CSV, JSON, and Avro
39. Livre Blanc de Tetra Pak : Industrie 4.0
40. Etude Gartner : How to Establish a Data-Driven Culture in the Digital Workplace
41. Etude Decisive Action : How Businesses Make Decisions and How They Could do it Better
42. Article Dataconomy : Supermarket Giant Tesco pioneers Big Data: Turning Customer Loyalty into Royalties
43. Article Harvard Business Review : Tesco's Downfall Is a Warning to Data-Driven Retailers
44. Livre de Bill Gates : Business at the Speed of Thought: Succeeding in the Digital Economy
45. Livre de Viktor Mayer-Schonberger et Kenneth Cukier : Big Data: A Revolution That Will Transform How We Live, Work, and Think
46. Video de SAS Engagement Leader Lisa Loftis : What Is Data Governance?
47. Etude de Seagate : The Digitization of the World From Edge to Core
48. Livre de Prashanth Southekal : Data for Business Performance: The Goal-Question-Metric (GQM) Model to Transform Business Data into an Enterprise Asset
49. Webinar TDWI de Fern Halper et David Stodder : What It Take to be Data-Driven
50. MOOC de Kelley OConnell : Culture of measurement and growth course
51. SlideShare présentation M6 : AB Testing chez M6Web
52. Etude de John R. Hauser et Gerald M. Katz : Metrics: You Are What You Measure!
53. Vidéo de Cameron Davies, McKinsey : What we demand, expect, and accept
54. Vidéo de Rob Casper, McKinsey : Data is the crown-jewel asset
55. Vidéo de Jeff Luhnow, McKinsey : Real people there to help
56. Livre de John Akred, Edd Wilder-James, Scott Kurth : What Modern Data Technology Makes Possible
57. Etude Deloitte : API economy - From systems to business services
58. Livre de Douglas B. Laney : Infonomics: How to Monetize, Manage, and Measure Information as an Asset for Competitive Advantage
59. Livre Blanc, GUSS : Comment choisir sa solution décisionnelle
60. Etude Business Application Research Center : BI Trend Monitor 2019
61. Livre de Nathan Marz et James Warren : Big Data Principles and best practices of scalable realtime data systems
62. MOOC Pluralsight : Recognizing the Need for Data Literacy
63. MOOC EDX : Microsoft Professional Program for Internet of Things
64. Ebook Cigref : Guide d'audit de la gouvernance du système d'information de l'entreprise numérique