



Y A-T-IL UNE VIE APRÈS LES POCs ?

RÉUSSIR L'INDUSTRIALISATION DU BIG DATA



AVEC LA PARTICIPATION DE :

Jeremy Harroch, Guillaume Perrin-Houdon, Emmanuel Manceau, Long Do Cao,
Gill Morisse, Matthieu Vautrot, Ysé Wanono, Stéphane Jankowski, Nicolas
Gibaud, Abdellah Kaid Gherbi, Youssef Benchekroun, Alberto Guggiola,
Vincent Dejouy, Jean-Matthieu Schertzer, Robin Lespes, Michèle Marchand.



Quantmetry
Data Science Consulting



criteo



SNCF



BAKER & MCKENZIE



SAFRAN
AEROSPACE · DEFENCE · SECURITY



bouygues
TELECOM

×

Y A-T-IL UNE VIE APRÈS LES POCs ?

RÉUSSIR L'INDUSTRIALISATION DU BIG DATA

×

S O M M A I R E

INTRO : Accélérer la création de valeur en industrialisant le Big Data et les analytics.....10

PARTIE I : S'organiser autour de la donnée.....16

I.1	Big Data : Du Lab à la Fab.....	16
I.1.1	Les projets Big Data, l'iceberg.....	16
I.1.2	Responsabilités de la chaîne data.....	16
I.2	Piloter ses projets.....	19
I.2.1	Estimer la complexité d'un projet.....	19
I.2.2	Qui finance quoi ?.....	20
I.2.3	Gouvernance des initiatives.....	21
I.2.4	Témoignages.....	22
I.3	Composition des équipes.....	25
I.3.1	L'équipe pour le POC.....	25
I.3.2	De nouveaux profils.....	25
I.3.3	Le gestionnaire de programme.....	27
I.3.4	Témoignages.....	27
I.4	Le règlement sur les données.....	29
I.4.1	Une nouvelle réglementation.....	29
I.4.2	Témoignages.....	29

PARTIE II : Conduite du changement.....33

II.1	Transformer les processus métiers.....	33
II.2	Témoignages.....	35
II.3	Interprétabilité des résultats.....	38
II.3.1	Quand privilégier l'interprétabilité à la performance ?.....	38
II.3.2	Trouver le bon compromis.....	39
II.4	Datavisualisation.....	40
II.4.1	Levier de la compréhension.....	40
II.4.2	La datavisualisation « user-centric ».....	41
II.4.3	Le choix des technologies.....	41
II.4.4	Témoignages.....	42

PARTIE III : Architecture IT.....44

III.1	Du POC au pilote.....	44
III.1.1	POC vs pilote, quelle(s) différence(s) ?.....	44
III.1.2	À quelles contraintes faut-il penser lors du passage en pilote ?.....	45

III.2	Les patterns d'architecture.....	48
III.2.1	Les typologies de besoins.....	48
III.2.2	Cloud.....	51
III.3	Edition de logiciel, comment choisir ?.....	53
III.3.1	A quel moment choisir ?.....	53
III.3.2	Qui choisir ?.....	53
III.3.3	Quels critères ?.....	56
III.4	Data lake.....	57
III.4.1	Pourquoi un data lake ?.....	57
III.4.2	De la base de données relationnelle au data lake.....	58
III.4.3	Réussir la mise en place de son data lake.....	59
III.4.4	Témoignages.....	60
PARTIE IV : Vie quotidienne d'un modèle.....		63
IV.1	Cycle de vie.....	63
IV.1.1	Un modèle, vision mathématique approchée de la réalité.....	63
IV.1.2	La vie d'un modèle « primitif ».....	64
IV.1.3	Témoignages.....	64
IV.1.4	Apprentissage actif.....	66
IV.2	La maîtrise de la performance du modèle de bout en bout....	67
IV.2.1	Attention aux métriques des POCs.....	67
IV.2.2	Transformer la métrique scientifique en métrique métier.....	67
IV.2.3	Apprentissage par renforcement.....	68
IV.2.4	A/B Testing.....	69
IV.2.5	Vers Data Ops.....	69
IV.3	Dérive des données.....	71
IV.3.1	Quelle validation mathématique ?.....	71
IV.3.2	Simulation des déformations des données.....	71
IV.3.3	Détection de valeurs aberrantes en entrée.....	72
IV.3.4	Intervalles de confiance.....	72
IV.3.5	Logiciels pour traitement automatique du Machine Learning.....	72
IV.4	Effets de rétroaction.....	73
IV.4.1	Pression de sélection.....	73
IV.4.2	Effet systémique.....	74
IV.4.3	Dégénération de la base d'apprentissage.....	75
CONCLUSION.....		77



Emmanuel Manceau,
Manager chez Quantmetry



Jeremy Harroch
Fondateur et PDG de Quantmetry

P R É F A C E

Poussée par la pression concurrentielle des nés globaux, la data est présente dans les médias et tout l'écosystème digital depuis maintenant 3 ans en France. Cette renaissance de l'exploitation intelligente de la donnée a été largement adoptée par tous les acteurs de l'économie : on transforme des sociétés pour qu'elles deviennent data-driven, le marketing devient data-centric, les DSI sont Big Data et les fonds de Capital Venture investissent dans les start-up de la data.

Au pied de cette immense ambition qu'est l'Intelligence Artificielle, nous nous sommes rendu compte que la data n'est pas aussi utilisée qu'on pourrait le croire et bien des initiatives en restent à l'état de POC (*Proof Of Concept*). L'engouement pour ce sujet pousse le marché à proposer des offres disproportionnées et certains acteurs à minimiser la difficulté de sa mise en application. Aussi il nous paraît nécessaire de donner aux personnes en charge de développer des programmes autour de la donnée, une vision claire et des lignes de force pour éviter les écueils et réussir l'industrialisation. C'est à l'issue de ce cheminement que sera tenue la promesse du retour sur investissement.

Sans algorithme, sans Machine Learning, nous passerions à côté de la révolution que nous vivons qui est celle de l'interaction avec la data, de la consommation de la data et de son interprétation. Pas de voiture sans conducteur, de robot conversationnel (*chatbot*), de système de transaction décentralisé (*blockchain*), de médecine connectée ou de modèle d'affaire fondé sur la tarification à l'utilisation (*pay as you drive*) sans mise en production de ces modèles mathématiques.

Ce travail est le fruit de la synthèse de dizaines de missions réalisées par Quantmetry, acteur incontournable du Big Data et de la data science en France, éclairée par les retours d'expérience et témoignages de la SNCF, Bouygues Telecom, Criteo, BackerMcKenzie et Safran que je remercie chaleureusement pour avoir accepté de partager ici leur savoir.

..... *Jeremy Harroch*

REMERCIEMENTS

La rédaction de ce livre blanc a été menée en s'appuyant sur les retours d'expérience des entreprises en pointe sur les sujets Big Data. Aussi, nous alternons parties explicatives et extraits d'interviews qui viennent illustrer le propos.

Nous tenons à remercier chaleureusement Bouygues Telecom, SNCF, Safran, Criteo et Baker & McKenzie pour avoir consacré du temps à ce livre blanc.

Le Data Lab de Bouygues Telecom

Bouygues Telecom dispose depuis 2013 d'une structure baptisée Data Lab chargée de la promotion des projets Big Data. Son rôle s'est étoffé l'année dernière en vertu de la décision stratégique de l'opérateur télécoms de faire du Big Data un vecteur de transformation de l'entreprise. En 2016, l'opérateur installe la rampe de lancement censée propulser l'analyse complexe de données vers une phase plus industrielle en envisageant notamment de se doter d'un lac de données.

La Big Data Fab de la SNCF

Opérationnelle depuis un an, la Fab Big Data est une entité rattachée à la direction digitale du groupe de transport ferroviaire. Son rôle consiste à fournir aux entités métiers du groupe les moyens de développer des projets d'analyse de données destinés à la prédiction des pannes, des incidents ou encore la projection de chiffre d'affaires. Techniquement, la Fab Big Data centralise les données de l'entreprise au sein d'un lac de données afin de faciliter leur manipulation et met à disposition des outils d'analyse aux métiers capables de les utiliser en autonomie.

Criteo

Tout le monde a rencontré ces bannières publicitaires ultra-ciblées qui vous suivent d'un site à l'autre, vous rappelant les articles repérés sur un site marchand. Derrière ce tour de passe-passe se cachent les technologies de l'ancienne start-up devenue leader mondial à succès Criteo, pionnière du retargeting publicitaire et introduite avec succès au Nasdaq. Son architecture informatique est à la pointe du domaine Big Data.

Safran

Safran s'est doté d'une entité à part entière afin de fédérer et de coordonner les actions du groupe dans le domaine du Big Data, et de faire de l'exploitation des données un nouveau levier de performance. Safran Analytics intervient en transverse de toutes les entités du groupe sur des sujets variés tels que les services clients, la prédiction des pannes ou encore la production.

Le département ITC de Baker & McKenzie

Baker & McKenzie est un des plus grands cabinets d'avocats d'affaires au monde avec près de 10 000 collaborateurs dont 4 100 avocats. En France, le cabinet est leader sur

le marché en droit des Technologies de l'Information et de la Communication, conseille des clients publics et privés sur toutes les questions relatives aux technologies, à la protection des données personnelles et aux télécommunications, et entretient d'étroites relations avec les autorités de régulation françaises et européennes.



SNCF, Big Data Fab



Marine Mizrahi,
Program Manager,
Direction du Digital



Bouygues Telecom,
Data Lab



Sylvain Goussot
Directeur Big Data
et Innovation



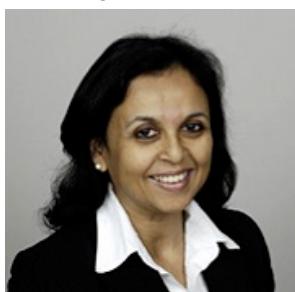
Criteo, R&D



Nicolas Leroux,
Responsable R&D (Photo)
Yann Schwartz,
Ingénieur Software

BAKER & MCKENZIE

Baker & McKenzie Paris
Département, Information,
Technologies, Communication



Denise Lebeau-Marianna
Partner en charge la pratique
« données personnelles »



SAFRAN
AEROSPACE · DEFENCE · SECURITY

Safran Analytics



Philippe Naim,
Directeur Stratégie
et Développement,
Safran Analytics



Julien Ricordeau,
Senior Data Scientist

INTRO : Accélérer la création de valeur en industrialisant le Big Data et les analytics

« Big Data », « analytics » ; ces termes rencontrent un succès grandissant dans les articles de la presse business, dans la bouche des dirigeants ou encore chez les opérationnels qui y voient un champ d'opportunités à explorer.

Après plusieurs années consacrées par les « **early adopters** » (ces entreprises qui ont commencé à mettre en œuvre des initiatives Big Data et analytics entre 2010 et 2016) à explorer ces opportunités et à les tester via la réalisation de **Proof Of Concept (POC)**, 2017 est une année charnière où les préoccupations se portent sur les questions d'industrialisation et d'accélération de la création de valeur de ce type d'initiatives.

DE QUELS ENJEUX DE CRÉATION DE VALEUR PARLE-T-ON ?

Le potentiel de création de valeur ne réside pas dans les données massives (Big Data) à elles seules, mais plutôt dans **la capacité à construire des analyses et modélisations avancées s'appuyant sur ces données** et dans la capacité à les mettre en œuvre rapidement dans le cadre de cas d'usage bien identifiés, intégrés de manière tangible aux processus opérationnels de l'entreprise.

Les initiatives « Big Data » reposent en effet sur un ensemble d'innovations et d'optimisations technologiques ayant ouvert un champ d'opportunités nouvelles pour les entreprises : stocker, traiter et exploiter des données plus volumineuses, plus variées, plus véloces. Ces innovations ont également rendu possible la construction d'analyses avancées et de modèles prédictifs s'appuyant sur ces données : on parle de « **Machine Learning** » ou d'**« analytics avancés »**.

Du point de vue business, les possibilités offertes par les « advanced analytics » sont très variées. L'enjeu pour les entreprises est de savoir identifier celles qui sont les plus pertinentes dans leur contexte et de les décliner opérationnellement pour en faire un levier de différenciation, voire un avantage compétitif. Nous identifions usuellement 4 grands domaines d'opportunités, illustrés ci-après avec quelques exemples :

- ✖ **optimisation de l'offre et de l'expérience client** : mieux comprendre les clients, bâtir des approches permettant un ciblage très précis pour optimiser la performance marketing ; personnaliser les offres et les interactions pour accroître la satisfaction client ;
- ✖ **amélioration de l'efficacité des processus opérationnels** : estimer la durée de vie du matériel et mieux prédire les pannes ; mieux piloter la qualité des procédés de production industriels ; prédire la demande pour mieux optimiser la supply chain ;

- ✖ **renforcement de la gestion des risques** : améliorer les processus de détection de la fraude ; raffermissement des solutions de cybersécurité ;
- ✖ **valorisation et monétisation de la donnée** : construire des offres de produits et services s'appuyant sur le patrimoine de données à disposition de l'entreprise.

INDUSTRIALISER LE BIG DATA ET LES ANALYTICS ?

Industrialiser le Big Data, c'est mettre l'entreprise en capacité de gérer et traiter de manière pérenne les données à sa disposition et de pouvoir injecter dans ses processus les résultats de traitements analytiques s'appuyant sur ces données. Ceci doit permettre de dégager des bénéfices tangibles, répétables à grande échelle et dans la durée.

En 2017, l'enjeu pour beaucoup d'entreprises est d'organiser le passage du « laboratoire à l'usine ». Concrètement, il s'agit de commencer à faire passer massivement les différents POCs réalisés par les **Data Labs** dans des environnements « bac à sable » à la « vie réelle », avec une utilisation opérationnelle des modèles prédictifs développés et un impact réel sur les processus métiers. Cela implique de mettre en place un ensemble de dispositifs permettant de :

- ✖ **structurer une stratégie et une feuille de route d'initiatives data**, alignée sur les enjeux de transformation métier et la stratégie de l'entreprise ;
- ✖ **identifier, capturer, stocker de manière consolidée et traiter les données massives** avec des approches et outils présentant un niveau d'automatisation adapté aux enjeux. (Par exemple : la notion de « traitements en temps réel » peut-être coûteuse et n'est pas toujours un impératif premier de la feuille de route data) ;
- ✖ **développer des modèles et solutions analytiques industrialisées** (scores, modèles prédictifs,...) et **savoir les intégrer aux processus opérationnels** (outils d'aide à la décision, automatisation de tâches, ...) ;
- ✖ **gérer et maintenir dans la durée le patrimoine de solutions analytiques** développées, avec une dynamique adaptée aux métiers qu'elles servent. (Par exemple : chez certains acteurs du web, un jour de retard dans la mise à jour de modèles prédictifs peut avoir un impact sur le chiffre d'affaires de l'ordre de quelques pourcents).

UN ENJEU STRATÉGIQUE

L'industrialisation du Big Data et des analytics est un enjeu stratégique, qu'il convient de gérer au bon niveau dans l'entreprise.

Tout d'abord, l'industrialisation est le levier qui permet de dégager la promesse de valeur tant attendue. En effet, les initiatives Big Data & analytics à plus forte valeur et potentiellement porteuses d'innovations de rupture sont celles qui reposent sur le Machine Learning et qui nécessitent le plus d'automatisation (mécanismes de scoring automatisés, algorithmes de détection prédictifs...). C'est en les industrialisant et en les faisant passer à l'échelle que l'entreprise pourra dégager des **gains significatifs et durables**.

Pour apporter toute la valeur recherchée, c'est également un sujet qui doit être pris en main à l'échelle de l'entreprise. Les enjeux associés résident tout particulièrement dans la mise en place des mécanismes techniques et des modes de gouvernance permettant de casser les silos organisationnels et de libérer l'utilisation de la donnée de manière transverse aux différentes entités métier.

L'industrialisation induit de fortes exigences en termes de qualité et de maîtrise d'exécution.

En effet, les exigences accrues en matière de gestion de la qualité et de la sécurité de la donnée nécessitent des dispositifs de gouvernance spécifiques.

La maîtrise des modèles prédictifs est également un enjeu de taille : les exemples ne manquent pas de situations où les algorithmes se sont « emballés » ou tout simplement « trompés », faisant peser sur l'entreprise des risques, d'image notamment, pouvant être significatifs.

La mise en production de solutions data industrielles suppose également de mettre en place un ensemble de dispositifs permettant de garantir leur bon niveau de performance technique : infrastructures technologiques robustes, SLAs, dispositifs de supervision...

Il est également important d'ajouter que l'excellence d'exécution ne se limite pas à des sujets techniques : il est en effet plus difficile de « poser la bonne question » que de savoir trouver l'algorithme qui permet d'y répondre. Les entreprises doivent donc également intégrer des compétences spécifiques de gestion de projet et de conseil afin de gérer efficacement les transformations induites par les projets data.

Enfin, la gestion industrialisée du Big Data induit également l'instruction de questions sensibles (éthique, protection des données, sécurité...) autour desquelles l'entreprise doit s'organiser et mettre en œuvre une transformation culturelle parfois profonde.

IL EST IMPORTANT DE RÉFLÉCHIR SUFFISAMMENT EN AMONT À UNE STRATÉGIE D'INDUSTRIALISATION

L'enjeu des démarches d'industrialisation est de permettre une accélération du rythme de création de valeur liée à la mise en œuvre des initiatives data : maximiser l'apport de valeur unitaire des initiatives, accélérer le rythme de production de nouvelles initiatives, diminuer leur « time-to-market » unitaire. Pour assurer cela, l'industrialisation doit être prise en compte très en amont.

Il ne suffit pas d'attendre que les POCs soient réussis pour réfléchir à la question de leur industrialisation. L'accélération et le passage à l'échelle nécessitent de prendre en compte les questions d'industrialisation tout au long de la chaîne de valeur des initiatives Big Data, avec un ensemble de questions à traiter à chacune des étapes :

- ✖ **Lors de l'identification et de la sélection des projets** de la feuille de route Big Data : comment choisir les initiatives qui seront, d'une part, porteuses du plus grand potentiel de valeur mais également celles qui contribueront à bâtir le socle de compétences, méthodologies et outils nécessaires à l'industrialisation et permettront ainsi une accélération et un passage à l'échelle dans l'entreprise ?
- ✖ **Réalisation des POCs** : comment profiter de chaque réalisation pour contribuer à la démarche d'industrialisation globale et permettre une mise en œuvre accélérée des prochaines initiatives sur des sujets ou des périmètres de données proches ? Comment anticiper les questions d'industrialisation de manière à optimiser le produit de sortie du POC et limiter l'effort nécessaire à son opérationnalisation (réécriture de code, gestion des versions d'environnements,...) ? Comment anticiper au mieux le passage à un mode opérationnel, où l'algorithme s'implante dans les processus métiers ?
- ✖ **Préparation du déploiement (phase pilote)** : comment gérer au mieux cette phase pour qu'elle permette de préparer une transition la plus efficace possible entre un concept et un objet déployé et opérationnel en production ?
- ✖ **Exécution opérationnelle et pilotage au quotidien des solutions analytiques déployées** : comment construire efficacement et progressivement, au rythme des nouvelles initiatives arrivant en production, le dispositif qui permettra de gérer le patrimoine de solutions analytiques de l'entreprise ?

Le CDO, prêt à attaquer l'industrialisation



L'INDUSTRIALISATION DU BIG DATA DOIT ÊTRE ORCHESTRÉE COMME UN PROJET DE TRANSFORMATION.

Lorsque la stratégie Big Data & analytics est posée, c'est-à-dire que les domaines prioritaires et opportunités associées ont été identifiés et que l'entreprise s'est organisée pour rassembler les premiers moyens pour les délivrer (feuille de route technologique, compétences, briques d'organisation), il est grand temps de penser « industrialisation ».

Quatre dimensions clés doivent être prises en compte et instruites de manière coordonnée dans le cadre d'une feuille de route globale, allant au-delà de la simple exécution du portefeuille d'initiatives identifiées :

- ✖ **Socle technologique** : mettre en place le socle nécessaire au passage à l'échelle des projets réalisés en POCs ; assurer l'interfaçage avec les SI et technologies existants ; assurer la montée en maturité rapide de l'organisation et des équipes sur ces nouvelles technologies ;
- ✖ **Organisation data** : mettre en place l'organisation, les compétences et modèles de delivery compatibles avec les exigences de montée en puissance (plus d'initiatives, plus fréquemment) ;
- ✖ **Gestion du patrimoine de données et de modèles prédictifs** : définir l'organisation en charge de la gestion du cycle de vie des modèles ; mettre en œuvre les dispositifs permettant de surveiller et de faire évoluer au fil du temps le « portefeuille de modèles » de l'entreprise ; mettre en place une gouvernance de la donnée adaptée au contexte et aux enjeux ;
- ✖ **Gestion du changement** : diffuser la culture de l'algorithme et de la donnée et sensibiliser sur les opportunités associées ; accompagner la prise en main de ces outils au service des processus opérationnels de l'entreprise.

La démarche d'industrialisation est donc une transformation qui doit orchestrer l'ensemble de ces dimensions, sur lesquelles il s'agit de monter en maturité progressivement et de manière coordonnée.

Il est également important d'appréhender l'ensemble de ces sujets avec une approche itérative et une logique d'amélioration continue (« learn by doing ») et enfin, ne pas oublier d'être pragmatique : « **savoir faire simple et qui fonctionne, avant de faire compliqué** ».

..... *Guillaume Perrin-Houdon*

PARTIE I : S'organiser autour de la donnée

I.1 | Big Data : Du Lab à la Fab

I.1.1 - LES PROJETS BIG DATA, L'ICEBERG

Les premiers temps du Big Data ont vu l'émergence de structures d'innovation souvent baptisées **Data Labs**. Ces entités transverses, parfois isolées d'un métier particulier (et de la DSI) ont pour mission **d'identifier des initiatives et de réaliser des expérimentations**. Le Data Lab est doté de ressources humaines et financières propres pour accompagner l'entreprise dans la réalisation des projets data. Isolé, soumis à des contraintes allégées (normes de sécurité allégées, procédures d'achat simplifiées), le Data Lab est devenu un terrain propice à l'innovation. Avec ces éléments favorables, le Data Lab a favorisé l'implantation d'une culture data au sein de l'entreprise.

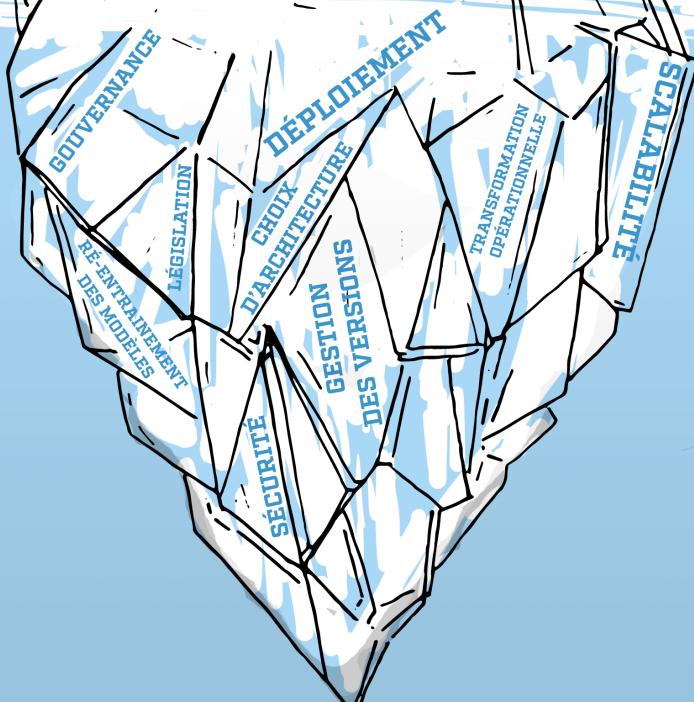
Pourtant, après quelques années de fonctionnement, un doute est apparu quant à l'efficacité de tels dispositifs. Face aux nouvelles attentes, les Data Labs réagissent en révisant leur organisation et leurs compétences. En effet, **industrialiser un prototype fait surgir des problématiques jusqu'alors immangées. La complexité du déploiement, l'impact sur les processus opérationnels ou encore la confrontation avec les données vivantes** font que les coûts augmentent fortement tandis que les niveaux de performance s'amoindrissent. Une autre difficulté encore plus profonde réside dans le choix des sujets, le potentiel de transformation des processus ou encore le déploiement à large échelle des projets. Face à des demandes multiples, le Data Lab a des difficultés à prioriser et n'obtient pas forcément de participation suffisante des directions métiers prescriptrices. Enfin, si le temps d'un POC toutes les contraintes de l'entreprise s'évanouissent, cette disparition n'est que temporaire et dès le pilote, celles-ci réapparaissent alors même que le Data Lab n'y est pas préparé : **gouvernance des données et mise en qualité, choix d'architecture, organisation des opérations, respect des normes et standards** (sécurité, directives CNIL, législateurs).

I.1.2 - RESPONSABILITÉS DE LA CHAÎNE DATA

Face à ces difficultés, nous observons une transformation du Data Lab traditionnel vers une Data Fab. Le changement de vocabulaire manifeste l'ampleur de la transformation : nous passons du laboratoire à la fabrique. Fabrique ou usine à score, production de modèles à grande échelle, le modèle Data Fab redessine les frontières organisationnelles et se positionne comme un acteur de l'entreprise dont le rôle est de délivrer des services fiables, générateurs de revenus ou d'améliorations de services.

POC

MODÈLE



INDUSTRIALISATION

Du POC à l'industrialisation, la Data Fab responsable sur toute la ligne :

La Data Fab intègre, dès le départ, la question de l'industrialisation et de l'exploitation de plateformes Big Data. Elle porte l'engagement de niveau de service vis-à-vis des autres entités en les accompagnant pendant la phase d'industrialisation et de déploiement. Elle est responsable du bon fonctionnement du Big Data, des flux d'alimentation et des modèles. Au même titre que le mouvement DevOps a aplani les frontières existant entre études et exploitation, **la Data Fab est responsable, non seulement des phases projet, mais aussi de la vie courante du modèle.**

Pour déplacer les lignes, détecter les sujets prometteurs, les Data Fab se créent des identités visuelles fortes, coordonnent des actions de communication interne et externe, et se dotent d'un ensemble de moyens pour s'interfacer avec les entités métiers. En s'appuyant davantage sur l'expertise terrain, **en confiant aux responsables métiers le rôle de product owner**, la Data Fab renforce sa capacité d'alignement avec les exigences de performance opérationnelle et accroît sa capacité à créer de la valeur.

Enfin, parce que cette entité ouvre la voie à une utilisation massive du patrimoine informationnel de l'entreprise, elle se voit souvent confier la mission d'engager des actions de long terme destinées à améliorer et favoriser la qualité de la donnée. **Elle pilote alors les actions préventives pour améliorer les données à la source plutôt qu'en correctif.** Elle peut à la fois exiger une meilleure saisie dans les outils mais surtout travailler sur l'ergonomie des applications sources afin de rendre la saisie plus simple et plus efficace.

Profiter à grande échelle de la valeur des données nécessite en fait de se confronter à une multitude de problèmes cachés pour lesquels se préparer est essentiel. Pour faire face et capter la nature intrinsèquement transverse des données, des entités autrefois isolées doivent se mettre à collaborer par l'entremise de la Data Fab.

..... *Emmanuel Manceau / Long Do Cao*

I.2 | Piloter ses projets

La réussite d'une initiative Big Data repose sur une estimation correcte des ressources à engager et sur la **maîtrise budgétaire** du projet. Rien n'est plus destructeur qu'un projet dont les échéances, le budget et le Reste À Faire (RAF) sont constamment revus à la hausse. Une industrialisation étant particulièrement onéreuse, trois questions se posent alors :

- 1.** Comment estimer le coût d'un projet ?
- 2.** Qui doit financer ?
- 3.** Comment identifier et prioriser un ensemble d'initiatives, selon quelles règles de gouvernance ?

I.2.1 - ESTIMER LA COMPLEXITÉ D'UN PROJET

Un projet data science implique souvent la création de processus nouveaux. Ces derniers s'accompagnent très généralement de besoins en budgets matériels ou humains. Plus ces nouveautés sont nombreuses, plus le projet peut s'avérer complexe et coûteux. Elles peuvent être classées en trois catégories.

Les coûts projets (développement, flux), socle (architecture, infrastructure matérielle et logicielle) et transformation de processus (conduite de changement, formation) viennent rapidement à l'esprit et sont bien identifiés dès le démarrage.

Le coût humain lié à la transformation digitale est généralement sous-estimé. Ceci peut inclure : la transformation des habitudes de travail d'un mode DSI à un mode digital (méthode agile, cycle court itératif), la transformation des politiques achats (passage d'une logique de volume, bas coût, fournisseurs généralistes à une logique de résultat, un fort niveau de compétences des ressources et fournisseurs spécialistes), la montée en compétence des équipes techniques sur des technologies peu matures, l'intégration de nouveaux rôles humains (comité éthique, data protection officer), ou encore la transformation des habitudes de travail d'une large population à travers une modification d'un ou plusieurs processus métiers.

Les coûts générés par le contrôle des résultats, c'est-à-dire la validation des résultats fournis par un automate, **la supervision des modèles**, ou encore **l'effet de mauvaises décisions** (usine, équipements arrêtés et envoyés à tort en maintenance, arrêt de production, déclenchement d'une pénalité, mauvais diagnostic médical) sont eux **souvent oubliés**.

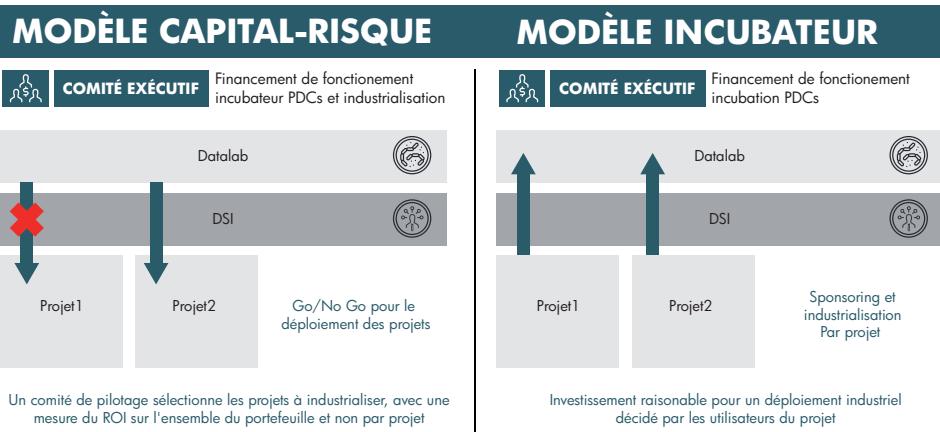
1.2.2 - QUI FINANCE QUOI ?

Par qui doivent être prises en charge ces nouvelles dépenses ? Si le budget d'un pilote est conséquent, mener plusieurs projets simultanément excède le budget de fonctionnement d'un Data Lab. Comment éviter de tarir les projets d'exploration ?

Une première approche est d'adopter une **attitude d'investisseur en capital-risque** et procéder à la sélection des projets en instituant des étapes de GO/NO-GO). A chaque étape (détection d'opportunité, POC, pilote, industrialisation), un comité de pilotage décide de la poursuite ou non d'un projet. **Les critères de GO/NO-GO doivent évidemment être décidés en amont.**

Par ailleurs, si l'on suit cette approche, **le retour sur investissement ne doit pas être calculé pour chacune des initiatives mais doit être consolidé au niveau du portefeuille des initiatives.** Typiquement, pour un portefeuille d'initiatives, la majorité d'entre elles n'aboutiront pas à un ROI positif, certaines présenteront un faible revenu, alors que seule une poignée incarnera un réel succès. Même s'ils peuvent générer de la frustration, ces projets avortés s'accompagnent généralement d'enseignements qui vont favoriser la réussite d'un projet ultérieur. Gardons en tête donc que ces tentatives sont une étape obligée, mais valorisante.

Une seconde solution est de **substituer un financement transverse par un financement branche pour les pilotes et l'industrialisation dans une logique d'incubateur.** En effet, les futurs utilisateurs sont les plus à même de décider du montant d'investissement raisonnable pour leur projet. La structure Data Lab joue un rôle d'incubation mais la responsabilité du déploiement (et donc du financement) incombe à l'entité utilisatrice. Elle conserve néanmoins la maîtrise budgétaire pour le financement de ses fonctions régaliennes (plateforme technique, gouvernance et mise en qualité des données, formation).



I.2.3 – GOUVERNANCE DES INITIATIVES

Indépendamment du mode de financement, la coordination du projet et le déploiement à large échelle suppose la mise en œuvre d'une gouvernance qui comprend la définition du processus d'alimentation, de priorisation puis de pilotage du portefeuille d'initiatives.

Alimenter le portefeuille d'initiatives

Comment identifier des sujets porteurs, comment mobiliser des acteurs et décisionnaires terrains pour explorer de nouvelles idées, déployer de nouveaux services ? Toutes les entreprises ont une maturité différente face aux processus d'innovation. Et devant l'éventail de possibilités offertes par la data science dans des sujets aussi variés que la supply chain, le marketing, la production industrielle ou encore la finance, il n'est pas aisément d'initier des projets aux contours peu précis. Pour faire face à cette difficulté, la SNCF a ainsi doté sa Data Fab de **gestionnaires de programmes** dédiés à l'animation du portefeuille d'initiatives auprès des métiers. Interlocuteurs uniques de la Data Fab, ils ont pour mission de créer des espaces de collaboration privilégiés avec des **responsables métiers**, eux-mêmes désignés pour être les « **data ambassadeurs** » auprès de leurs entités. Ce processus **d'alimentation bottom up**, propre aux grands groupes, pourra être remplacé par des **dispositifs top down** dès lors que les contours métiers sont plus restreints, les enjeux plus évidents. Il n'en souligne pas moins le besoin de s'appuyer sur les experts métiers pour se focaliser sur les initiatives les plus prometteuses.

Prioriser les initiatives

Une fois les sujets identifiés, il s'agit de définir des règles de priorisation des projets entre eux afin de limiter les conflits et les déceptions des parties prenantes des projets non retenus. Prioriser une initiative data science n'est pas un exercice fondamentalement différent des exercices de priorisation classiques. La mise en place d'instances de gouvernance assurant un sponsorship fort des initiatives retenues constitue un socle incontournable. Concernant les critères de sélection, **le ROI, la capacité d'industrialisation, les gains de productivité ou encore l'alignement avec la stratégie globale seront toujours mis en regard du coût, des difficultés techniques ou organisationnelles**. Nous distinguons quatre critères spécifiques aux activités de data science :

- ✖ **la disponibilité des données** : c'est un critère déterminant, l'absence de données fiables, exploitables ou facilement disponibles constitue un obstacle insurmontable en data science ;
- ✖ **la montée en compétence des équipes de data science** : un projet peut s'avérer être une excellente opportunité de monter en compétence sur des périmètres métiers spécifiques, de tester des méthodologies et d'ouvrir la voie à des initiatives nettement plus prometteuses ;

- ✖ **l'opportunité d'embarquement des métiers :** un projet peut permettre à la Data Fab de commencer à construire un relationnel avec un métier, de faire la preuve concrète de son potentiel ;
- ✖ **les exigences réglementaires :** le projet de règlement européen sur les données est ambitieux et aura des impacts significatifs sur la faisabilité des projets.

Piloter les initiatives

Une fois un ensemble d'initiatives lancées, il s'agit désormais de contrôler l'avancement des projets, d'éventuelles adhésions et points de blocage. Là encore, des espaces de collaboration entre les projects owners (métiers), l'entité data et l'ensemble des entités impliquées (IT notamment) permettront d'assurer la gouvernance opérationnelle efficace d'activités data science qui ont vocation à s'insérer dans un système d'information et de process métiers en constante évolution.

La gouvernance des initiatives constitue ainsi le jalon indispensable pour créer les conditions favorisant l'émergence puis le portage de projet jusqu'à l'industrialisation.

..... *Emmanuel Manceau / Long Do Cao / Gill Morisse*

I.2.4 - TÉMOIGNAGES

Bouygues Telecom

Quels changements ce passage à un mode industriel implique-t-il en matière de gestion des données ?

Nous n'avons évidemment pas attendu l'arrivée du Big Data pour travailler sur le sujet de la gouvernance de la donnée. Mais la création d'un lac de données nous a amenés naturellement à y réfléchir davantage. L'organisation choisie réunit aujourd'hui l'informatique, propriétaire des données de relation client/métier, et le réseau qui fournit une grande partie des informations d'un opérateur de télécommunications. S'y ajoutent trois piliers de la mise en conformité : la sécurité et le juridique, bien entendu ; et l'éthique qu'il me tenait à cœur d'introduire. Pour répondre aux préoccupations de nos clients quant à l'utilisation qui est faite de leurs données, nous avons en effet une responsabilité de pédagogie et de communication.

SNCF

Comment l'entité Big Data sélectionne-t-elle les projets à mener ?

La Fab Big Data n'est pas une organisation hors-sol. Nous pourrions développer plein d'idées dans notre coin et identifier ensuite si elles peuvent trouver des débouchés. Ce

n'est pas notre philosophie. Notre rôle consiste essentiellement à accompagner et pousser les projets. En clair, nous ne portons pas de projets à la place des métiers. Aucun chef de projet ne figure d'ailleurs au sein de la structure. Toutefois, nos équipes sont engagées dans un dialogue permanent avec les entités concernées. Nous mettons la main au portefeuille, notamment en phase exploratoire où le retour sur investissement est incertain. Néanmoins, de notre point de vue, le principal financeur a intérêt à être celui qui porte l'enjeu et le projet.

Dans un groupe d'une telle dimension, n'est-il pas compliqué de dénicher les projets pertinents que vous pourriez soutenir ?

Nous présentons régulièrement notre structure aux différentes entités de la SNCF. Toutefois, le fait d'avoir engrangé de premiers succès et d'être une des composantes du programme digital de la SNCF contribue à nous faire connaître. Cela explique sans doute que, pour l'instant, les projets viennent à nous assez naturellement. La Fab Big Data tourne aujourd'hui à plein régime. Elle emploie vingt-cinq personnes à temps plein pour une vingtaine de projets. Avant d'être lancé, un projet doit cependant remplir plusieurs conditions essentielles : l'adéquation à une attente réelle de la part des utilisateurs finaux, ainsi qu'un retour sur investissement potentiel. Sans oublier évidemment la capacité à unir nos compétences et travailler avec la direction métier concernée pour bâtir ensemble une solution. Sur ce point, nous nous rendons compte qu'il y a parfois des sur-attentes, que le « storytelling » ambiant autour du Big Data contribue à alimenter.

Comment mesurez-vous ce « gain potentiel ? » Avez-vous des métriques pour mesurer l'intérêt, les gains potentiels d'un projet ?

Oui, nous définissons des métriques pour chaque projet. Toutefois, elles peuvent évoluer au fil du projet : il arrive régulièrement que la conduite du projet permette justement d'avoir des évaluations plus précises des métriques utilisées. Sur un projet de maintenance prédictive, par exemple, nous nous sommes rendu compte que le nombre de minutes perdues par an, qui est une des métriques principales (car cela a un impact fort sur les gains de l'entreprise et la régularité pour les clients), était complètement sous-estimé ! En effet, la description du phénomène n'était pas complète : le fait de conduire le projet nous a permis d'avoir une meilleure métrique. Mais on essaye néanmoins toujours d'avoir des métriques dès le départ. Nous ne travaillons pas sur un projet si nous n'avons pas au moins un KPI sur lequel nous pouvons justifier le fait de mettre des moyens.

Quel est votre mode de financement ? La Big Data Fab a-t-elle un budget qu'elle répartit entre les différents projets ? Travaillez-vous plutôt en mode service ?

Comme nous cherchons l'industrialisation, nous voulons que le métier porteur soit pleinement responsable. Il nous semble donc important que le porteur du projet finance. Par ailleurs, les projets Big Data ne sont pas des projets très chers par rapport aux projets

SI en général. Jusqu'à maintenant, nous n'avons pas eu beaucoup de mal à trouver les financements nécessaires.

En revanche, sur les projets où il existe une incertitude sur la valeur qui sera créée, nous pouvons mettre la main à la pâte sur une première phase exploratoire. Tant que le futur ROI n'est pas prouvé, mais que nous y croyons, nous pouvons participer financièrement pour pousser le projet. Pour aller plus loin, il faut que le métier soit impliqué financièrement car sinon nous prenons un risque sur une implication constante jusqu'au déploiement. D'autant que certaines phases projet sont souvent riches en frustrations et sources de démotivation.

Safran

Comment Safran Analytics sélectionne-t-elle les projets dont elle a la responsabilité ?

Qu'il vienne d'une approche top-down, bottom-up ou d'un benchmark des cas d'usage, il faut que le problème remonté impacte fortement le processus de création de valeur d'une entité pour être sélectionné. Tout nouveau projet est considéré comme une opportunité, et passe par une phase de qualification basée sur un ensemble de critères tels que la qualité des données, leur richesse, l'implication du sponsor, ou encore les ressources nécessaires. Nous nous assurons que les conditions de réalisation des travaux sont réunies et que l'initiative sera un levier de transformation pour Safran.

Quels sont les moyens que vous avez trouvés efficaces pour convaincre les autres entités de l'apport de la donnée ?

Nous essayons d'être sur place en aidant sur des cas réels et nous accordons beaucoup d'importance aux retours d'expérience des équipes métiers. Ce travail collectif permet aux experts de constater assez vite les limites des modèles, ce qui fait tomber la peur du remplacement par les algorithmes. La simplification de l'accès à des données propres faisant la synthèse des différents SI est aussi une valeur ajoutée en soi, palpable par les métiers.

Par ailleurs, le prototypage sur données chaudes a une valeur de preuve beaucoup plus importante. Ce qui apparaît a posteriori comme net est au quotidien beaucoup plus complexe et les équipes apprécient l'éclairage que nous leur apportons sur l'analyse de données chaudes.

I.3 | Composition des équipes

I.3.1 – L'ÉQUIPE POUR LE POC

De sa conception à sa maintenance en passant par sa réalisation, tout projet doit observer un équilibre sain et évolutif entre compétences métiers, techniques, et organisationnelles. Les projets Big Data ne dérogent pas à cette règle.

Historiquement, la réalisation de projets Big Data a très rapidement mis l'accent sur la nécessité d'investir fortement sur des compétences techniques pour s'approprier les nouvelles technologies et techniques du Big Data. **C'était l'époque du « data scientist roi » et de la course au recrutement. Depuis les entreprises se sont rendu compte que le data scientist ne porterait pas seul les projets dans leurs ensembles et qu'il fallait une véritable équipe pour réussir.**



I.3.2 – DE NOUVEAUX PROFILS

Dans le cadre d'un POC, une équipe de 2 ou 3 data scientists expérimentés et soutenus par un sachant métier était suffisante. Mais pour un pilote, réussir le déploiement en conditions réelles, évaluer l'impact sur les processus opérationnels et les transformer, réussir à travailler en transversal (inter-métiers, inter-applications) ou encore bâtir des systèmes disponibles et alimentés en données de qualité ne font pas partie du bagage du data scientist. Il faut donc renforcer l'équipe avec les profils listés ci-dessous.

Des architectes IT vont pouvoir répondre aux questions types :

- ✖ Quelle architecture (en fonction du SI legacy) sera la mieux adaptée pour :
 - Exposer les résultats aux utilisateurs (sécurité, scalabilité) ?
 - Supporter les rafraîchissements des données ?

Des experts ou sachants métiers aux questions :

- ✖ Quels sont précisément les résultats qui seront utiles pour l'exploitation opérationnelle quotidienne ?
- ✖ Quelles règles appliquer en cas d'évolution sur la qualité de données (données manquantes, valeurs aberrantes) ?
- ✖ Quelle décision prendre si le niveau de confiance d'un résultat n'est pas satisfaisant ?

Des ressources internes ou consultants en organisation pourront répondre aux questions comme :

- ✖ Quels sont les impacts d'une industrialisation sur l'organisation :
 - Des activités (évolution voire refonte des processus de ventes, marketing, industriels...)
 - Des équipes associées aux activités impactées
 - L'application supprime-t-elle du travail effectué jusqu'alors manuellement ?
 - Des ressources auront-elles besoin d'être formées ?
- ✖ Qui doit et comment assurer la maintenance de l'application en production ?

I.3.3 – LE GESTIONNAIRE DE PROGRAMME

L'utilisation efficace de toutes ces ressources nécessite une gouvernance et un management de projet à même de lever tous les obstacles liés aux projets transversaux ayant un fort impact opérationnel. C'est ici qu'apparaît **le nouveau héros du Big Data : le directeur de programme**. Ce profil déjà connu pour les grands projets SI fait son apparition dans les Data Labs qui se transforment sous son impulsion en Data Fabs. Il s'agit de la personne qui prend en charge l'industrialisation du projet et porte l'engagement de l'entreprise.

Une activation intelligente de ces différents profils complémentaires au bon moment est clé pour la réussite de l'industrialisation du projet.

..... *Matthieu Vautrot*

I.3.4 – TÉMOIGNAGES

SNCF Digital

Quels sont les besoins que vous avez aujourd'hui sur ce sujet ? Quels sont les profils que vous recherchez ?

Nous avons trois grands pôles dans la Fab Big Data:

- ✖ Un pôle Data Science, qui réalise les prestations de data science et concentre la connaissance sur la donnée. Nous avons démarré en n'utilisant que des prestataires, puis nous avons commencé à recruter pour, assez rapidement, constituer une équipe mixte : 4-5 internes, complétés par des externes. Le fait de faire appel à des externes nous permet de benchmarker nos approches et de toujours challenger et améliorer nos façons de faire. De plus, sur des sujets pointus, il peut être plus pertinent de faire appel à des cabinets experts.
- ✖ Un pôle IT, qui réalise les prestations techniques et concentre les compétences d'architecture, de développement, d'exploitation et de support. Il s'agit du pôle le plus important en termes d'effectifs. Aujourd'hui, l'activité du pôle IT est réalisée par VSCT, DSI de Voyages-SNCF.com, elle-même filiale de SNCF, avec déjà une bonne expérience du Big Data pour ses besoins propres.
- ✖ Un pôle Programme, qui pilote l'activité globale de la Fab, fait le lien entre les besoins des métiers et les ressources de la Fab, assure le cadrage et le suivi des projets.

Nous avons aujourd'hui des difficultés de recrutement pour tous les profils, sur l'ensemble

de ces trois pôles. La difficulté ne porte pas que sur les compétences pointues, même si, par exemple, le recrutement d'administrateurs système ou de data scientists est particulièrement ardu. Les profils plus généralistes de responsables de programme présentant un équilibre entre les compétences relationnelles, la connaissance métier et l'appétence pour la technique sont également difficiles à trouver.

Safran

Quels profils constituent votre équipe et quels sont ceux que vous recherchez ?

Notre structure d'environ 50 personnes regroupe une dizaine de data scientists, une dizaine de data engineers et de software engineers, sept chefs de projets, deux ingénieurs cogniticiens, un UX designer, et des responsables de la communauté, de la transformation et de l'innovation. Outre le pilotage opérationnel, les chefs de projets Safran Analytics ont pour priorité l'accompagnement des métiers et la transformation des processus. Leur objectif est que nos produits soient effectivement adoptés. Il nous est difficile de trouver ces profils car nos data scientists sont souvent de jeunes diplômés et n'ont pas encore l'expérience d'une transformation organisationnelle. Par ailleurs, nous avons fait le choix de l'open source notamment parce qu'il est plus facile de trouver des profils spécialisés sur ces technologies.

I.4 | Le Règlement sur les données

I.4.1 – UNE NOUVELLE RÉGLEMENTATION

Nous avons choisi de nous concentrer sur l'évolution majeure qui va accompagner la réglementation sur les données personnelles produites par l'Union européenne.

Applicable en 2018 et concernant toutes les entreprises, cette réglementation apporte des éléments structurants qu'il faut intégrer dès maintenant dans sa feuille de route d'industrialisation.

I.4.2 – TÉMOIGNAGE

Baker & McKenzie

Le Parlement européen a adopté le 14 avril le règlement européen sur la protection des données (le «Règlement»), applicable en 2018 dont l'impact sur les entreprises sera majeur. Pensez-vous qu'elles soient en ordre de marche pour s'y conformer ?

Les grands groupes sont déjà en ordre de marche et ont pris conscience de l'enjeu. La situation est plus délicate dans les écosystèmes de données partagées (marketing, annonceurs, agences digitales, solutions de DMP) car jusqu'ici les responsabilités étaient diluées. Pour se mettre en conformité, ces écosystèmes devront définir clairement la chaîne de responsabilités. Le bénéficiaire du traitement ne peut s'exonérer de sa responsabilité, même si les données personnelles utilisées ne sont pas stockées chez lui. Chacun doit clarifier son statut dans la chaîne d'acteurs.

A quelles entreprises s'adresse le Règlement ?

A toute entreprise ayant un établissement en Europe et exploitant des données personnelles, mais aussi des entreprises non européennes traitant des données personnelles de résidents européens dans le cadre de leur activité, entre autres pour leur offrir des produits ou services pour assurer leur suivi, notamment comportemental, sur le territoire européen. Autrement dit, il n'y aura plus de distorsion de concurrence ou de limitations aux usages pour les entreprises européennes. Le Règlement s'appliquera à tous.

Comment ce Règlement s'intègre-t-il dans le cadre existant ?

Il sera d'application directe (avec la possibilité d'être complété par des dispositions de droit national) et vise à renforcer les droits des personnes concernées ainsi que les obligations tant du responsable de traitement (société à l'origine du traitement) que du sous-

traitant (prestataire de service agissant pour le compte du responsable de traitement). Il vise à apporter une meilleure protection en veillant à ce qu'il s'impose à tous ceux qui traitent les données de résidents européens de la même manière.

Comment l'autorité de contrôle va-t-elle s'assurer du respect du Règlement ?

Le Règlement comporte au cœur de son dispositif l'obligation d'« accountability » (rendre compte). L'entreprise sera donc responsable de documenter de façon claire et précise les obligations qui lui incombent en vertu du Règlement (notamment par une information adéquate des personnes dont les données sont traitées, la tenue d'un registre des traitements, des procédures visant à mettre en place le « privacy by design » et « privacy by default » afin d'intégrer la protection des données dans la conception des traitements et des technologies utilisées, etc.). L'autorité de contrôle peut être amenée à travailler avec les autres autorités (notamment, en droit de la consommation ou droit de la concurrence) pour s'échanger des informations.

De plus, les sanctions prévues par le Règlement sont fortement dissuasives. Elles peuvent atteindre 20 millions d'euros ou 4 % du chiffre d'affaires mondial de l'entreprise fautive (le montant le plus élevé étant retenu) en cas d'infractions considérées comme graves. La Commission européenne a récemment démontré sa détermination à propos de l'évasion fiscale et compte bien adopter la même fermeté pour l'application du Règlement.

Nous sommes conscients qu'aujourd'hui un algorithme ne peut par exemple pas refuser automatiquement une demande de crédit. Quel est le périmètre exact du nouveau règlement dans l'automatisation et la transparence des algorithmes (justification des décisions, publication en open source, explication des variables, profilage affectant de manière significative) ? Cela veut-il dire qu'il faudra motiver un refus de crédit et documenter les règles de calcul d'un score ?

Le Règlement prévoit la possibilité de ne pas se soumettre à un traitement automatisé donnant lieu à du profilage, sauf dans certains cas. Si un tel traitement est mis en place, un certain nombre de garanties doivent être apportées (notamment une information claire sur la logique qui sous-tend l'algorithme utilisé, les conséquences à l'égard de la personne concernée, la possibilité de constater le résultat et de demander une intervention humaine etc.).

Le Règlement met en évidence le concept de privacy by design et privacy by default. Sans entrer dans les détails techniques de l'implémentation, quel impact cela aura-t-il sur les data lakes et plus largement les architectures analytics réunissant un grand volume de données issues de multiples sources ? Comment valider auprès de l'autorité la prise en compte de ces exigences ?

Ces concepts de privacy by design et by default sont nouveaux et auront un impact assez fort. Ces principes visent à ce qu'un traitement soit conforme au Règlement dès sa conception. Intégrer des mesures techniques et organisationnelles telles que la pseudonymisation permet de répondre de manière effective aux critères réglementaires (minimisation de la donnée). Le critère fondamental est de minimiser la donnée et de limiter au maximum la collecte et l'accès à la donnée. Des outils informatiques estampillés « privacy by design » devraient donc de plus en plus faire la différence.

Le Règlement prévoit une plus forte responsabilité pour les sous-traitants et propose de mettre en place un code de bonne conduite. Que contient-il ? Qui certifie ? Est-ce obligatoire ?

Effectivement, le Règlement prévoit une responsabilité des sous-traitants, ce qui n'était pas le cas jusqu'ici. Ils seront conjointement responsables avec le responsable de traitement à l'égard de la personne concernée. C'est un changement majeur pour toutes les entreprises agissant comme prestataires.

Concernant les codes de conduite, le respect d'un code de bonne conduite, approuvé selon les conditions du Règlement, sera considéré comme un élément permettant de démontrer la conformité. Certains codes sectoriels existant actuellement pourraient faire l'objet d'une approbation à cette fin.

Est-ce que dans une étape de R&D, i.e. démarche exploratoire et « interne », la législation est plus souple sur l'utilisation des données personnelles ?

Les traitements à des fins de recherche scientifique, historique ou à des fins statistiques peuvent être soumis à des dérogations aux droits prévus par le Règlement, par les lois nationales des Etats membres, dans la mesure où les droits prévus par le Règlement peuvent rendre impossible ou entraver la réalisation des finalités poursuivies...

Je suis employeur et j'effectue des recherches sur des profils LinkedIn publics. Ai-je le droit d'utiliser ces informations personnelles dans mes traitements ?

Bien que cette information soit publique et visible de tous, elle ne peut être collectée à l'insu de la personne concernée. Seule une information et demande explicite de l'employeur au salarié pourrait permettre le traitement (mais la personne peut aussi s'y opposer) à condition qu'un tel traitement soit justifié et proportionnel à la finalité recherchée, ce qui est discutable.

Un nouveau rôle se dessine, celui de Data Protection Officer. Quel est son rôle ? De qui est-il redevable ?

Cette personne aura la responsabilité d'aider l'entreprise à se mettre en conformité avec le Règlement et tiendra à jour le registre des traitements. Cette fonction est effectivement exposée (car la personne pourra émettre un avis négatif sur des traitements réalisés par l'entreprise) et idéalement doit reporter à un membre de la direction. Le choix du DPO devra prendre en considération les conflits d'intérêts (département compliance et risques, RSE, direction de l'audit). Il s'agit d'une fonction à fort niveau de responsabilité et ayant une vision d'ensemble sur les usages des données personnelles dans une entreprise. Qui plus est, il aura un rôle de conseil et d'accompagnement auprès des différents départements fonctionnels (RH, Marketing, IT etc.).

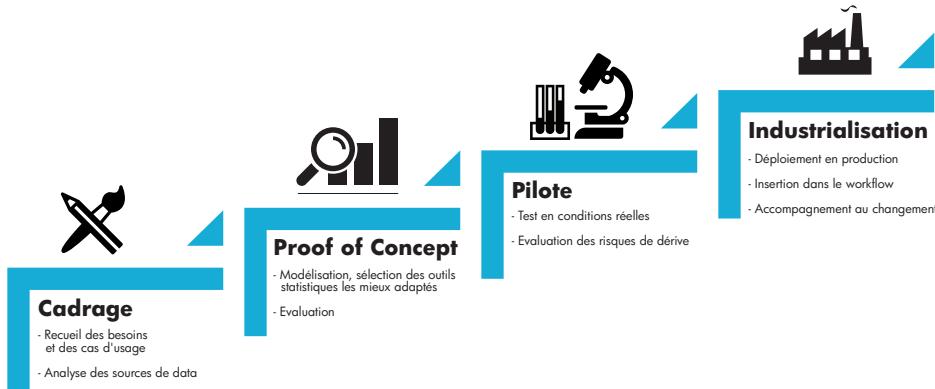
PARTIE II : CONDUITE DU CHANGEMENT

II. 1 | Transformer les processus métiers

Sans appropriation ni transformation des processus métiers, pas de gain de productivité ni nouveaux services. C'est dire si l'enjeu de transformation de processus métiers est essentiel dans la réussite d'une industrialisation.

Les projets Big Data ont deux types d'impacts sur les processus opérationnels :

- ✖ Ils améliorent les capacités d'aide à la décision des opérationnels.
- ✖ Ils accélèrent un processus métier via l'automatisation des tâches.
Ces projets ne vont donc pas sans soulever des questions et des craintes pertinentes des futurs utilisateurs.



- ✖ Comment interpréter les résultats d'un modèle ?
- ✖ Comment repenser le processus opérationnel ?
- ✖ Le Big Data ne va-t-il pas remplacer mon travail ?

Aussi, au vu de ces éléments, la refonte des processus et la **participation active des utilisateurs conditionnent le passage de la bonne idée à la réalité concrète.**

Certaines tendances se dégagent sur les meilleures formes que peut prendre cette démarche, comme par exemple l'organisation de présentations ouvertes à tous et destinées

à démystifier le Big Data, que ce soit pour lever les craintes ou pour tempérer les attentes excessives.

La présence d'un **sponsor fort du projet** constitue un des facteurs essentiels à une conduite du changement réussie. Il doit être suffisamment haut placé dans la hiérarchie afin d'être capable de mobiliser les différents acteurs de manière transversale, de les convaincre pour susciter leur adhésion, enfin d'actionner les leviers indispensables à la refonte d'un ou plusieurs processus opérationnels. Il est le premier sponsor de la démarche et favorise l'implication de tous. Il permet d'instaurer un climat de confiance entre les data scientists et les autres services impliqués dans la démarche Big Data.

La désignation d'un **responsable de processus (process owner)** travaillant en lien avec le chef du projet est un autre élément clé de la démarche. Ces responsables de processus existent déjà dans l'entreprise, souvent en lien avec une démarche qualité de type SixSigma ou Réglementaire. Ces interlocuteurs doivent être désignés dès la phase de POC et sont directement associés aux décisions de go-nogo en phase pilote et industrialisation.

En termes de cycle projet, il est essentiel à nos yeux que la **phase de cadrage du pilote incorpore une étude d'impact sur la refonte de processus**. Plutôt qu'aborder la conduite de changement en fin de projet, il convient de travailler au plus tôt avec les parties prenantes à repenser la place des acteurs dans le processus. Sans cet effort, le risque de rejet est réel et les pratiques ne seront pas transformées.

En phase d'industrialisation, le déploiement auprès des utilisateurs doit être soigneusement préparé, notamment parce qu'une aide à la décision mal comprise peut conduire à de mauvaises décisions. Sessions de formations, guides utilisateurs, participation à des démonstrations, implication des référents dans l'équipe projet, tous les moyens sont bons pour embarquer les utilisateurs.

Ce climat de confiance, ce partage de la culture Data au sein de l'entreprise, nécessite néanmoins que pour chaque initiative les objectifs de transformation des processus métier soient clairement posés dès les phases de cadrage initial. Il est important de bien définir en amont la stratégie à adopter, d'associer au plus tôt les parties prenantes à la transformation des processus et s'assurer que les compétences des acteurs sauront suivre les évolutions apportées par le Big Data.

.....Ysé Wanono

II.2 | Témoignages

Bouygues Telecom

Quelle est votre approche pour passer de l'expérimentation à la mise en production d'un modèle ?

La phase de détermination de faisabilité (POC ou proof of concept) permet de valider le concept, l'approche scientifique et la qualité des données nécessaires à un bon résultat. En cas de validation, elle est rapidement suivie d'un pilote en condition de production. Pilote et industrialisation sont intimement liés. Au cours de cette étape, nous faisons tourner en parallèle l'ancien système basé sur des données plus « déterministes » et le nouveau système afin de tester la performance de l'algorithme. C'est ce que nous avons réalisé par exemple dans le domaine de la détection de fraude.

SNCF Digital

Comment qualifiez-vous votre relation avec les différentes directions ?

Le dialogue est assez naturel sur la partie modélisation des données car un « data scientist » a intrinsèquement besoin de comprendre les données manipulées et de tester ses analyses. L'interaction y est obligatoire. Elle est plus compliquée sur les aspects très IT où il y a souvent des difficultés de compréhension de part et d'autre. La Fab Big Data n'a pas été construite comme un centre de services, nous construisons nos solutions en même temps que nous accompagnons des projets, le tout dans un écosystème complexe. Cela signifie que pour un chef de projet métier, un chantier Big Data nécessite souvent plus d'investissement dans la compréhension des enjeux techniques que ce à quoi il est habitué. Autant la modélisation des données génère du fantasme et de l'excitation, autant la partie technique – très forte en phase pilote, fait essentiellement émerger des contraintes et génère des tensions. Les projets suivent toujours grossièrement le même cycle : à la phase de lune de miel du POC succède une phase de désamour un peu plus difficile dans les phases d'industrialisation.

Safran

Dans un groupe d'une telle dimension, comment fait-on pour identifier des projets avec un fort potentiel de transformation ?

Le Big Data ne crée pas de valeur sans l'appropriation des modèles par les équipes et la transformation des modes de travail. Le problème réside donc essentiellement dans l'identification des bons cas d'usage et dans l'accompagnement des métiers pour une véritable transformation de notre organisation.

Pouvoir pivoter en cours d'étude est également essentiel pour nous car la première piste

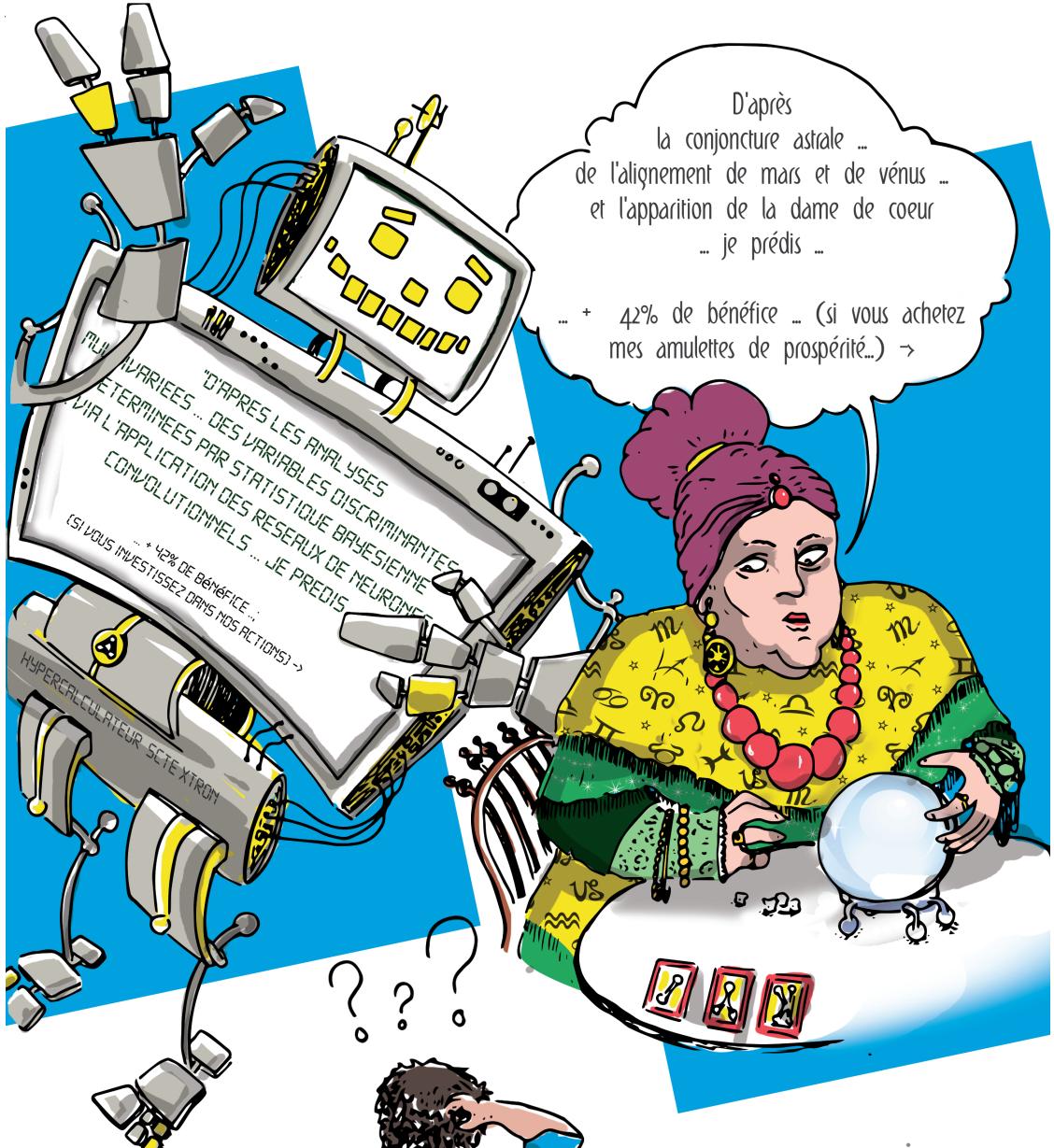
peut ne pas être la bonne : la valeur apparaît souvent là où nous ne l'attendions pas car la mise en relation de données hétérogènes met en relief de nouveaux aspects des projets.

Il est donc nécessaire que les experts soient fortement impliqués dans les projets pour déceler quelle méthode, quels indicateurs, quels outils peuvent apporter de la valeur. Ce travail est collectif car les opérationnels n'ont pas forcément d'idées précises sur ce qu'il est possible de faire avec les données.

Identifiez-vous des freins à l'industrialisation de projets de data science au sein de votre entreprise ?

La data science est une approche nouvelle pour certains secteurs et la convergence entre utilisateurs et data scientists est primordiale. La forte culture d'ingénierie de Safran est donc un atout pour faciliter la mise en place de solutions de data science.

L'industrialisation de ces solutions reste cependant une étape particulièrement complexe en particulier quand on considère les problématiques techniques pour lesquelles plusieurs questions restent en suspens. Quel niveau de développement logiciel devons-nous considérer pour nos produits ? Comment allons-nous industrialiser des technologies open source ? Qui assurera la mise à jour des modèles, la maintenance et la pérennité des solutions ? Les termes de l'industrialisation seront certainement définis au cas par cas selon le degré d'attente des clients et leur capacité à monter en compétence pour être autonomes sur la mise en œuvre des modèles.



HEU.....
VOUS ÊTES
^ SÛRS !?

II.3 | Interprétabilité des résultats

Les modèles d'apprentissage automatique sont souvent considérés au sein des entreprises comme des algorithmes performants, mais souvent difficiles à comprendre. Cette couche d'abstraction provient de la nature même des modèles de Machine Learning, comme le gradient boosting ou les réseaux de neurones. Ces modèles « boîtes noires » peuvent cependant laisser leur place à des modèles « boîtes blanches » privilégiant une meilleure interprétabilité.

II.3.1 – QUAND PRIVILÉGIER L'INTERPRÉTABILITÉ À LA PERFORMANCE ?

Un modèle d'apprentissage automatique n'est pas un système isolé. De fait, il évolue nécessairement dans un contexte opérationnel. S'il est initialement construit dans le but d'atteindre un certain niveau de performance, sa finalité est d'être utilisé par les acteurs métiers. Ainsi, dans le cadre d'une aide à la décision, il est alors essentiel que le modèle puisse être interprétable, permettant ainsi aux utilisateurs de pouvoir prendre des décisions basées sur la donnée. Il existe même des cas où la performance n'est pas le but in fine. Par exemple, dans un contexte de ressources humaines pour la prédition de l'absentéisme, on va privilégier la recherche des facteurs à l'identification des individus. Trouver et comprendre ces facteurs est une réelle valeur ajoutée puisqu'elle peut résulter en des actions correctives.

Par ailleurs, **la capacité à interpréter le modèle rend la communication et la compréhension de ses résultats bien plus accessible à tous**. Citons en exemple un acteur du transport ferroviaire qui, dans le cadre de la prédition des pannes sur les trains, a eu besoin d'un bon niveau d'interprétabilité pour que les techniciens puissent comprendre les origines des problèmes sur lesquels ils interviennent. En effet les techniciens qui travaillent en maintenance ferroviaire ont besoin d'informations sur l'origine de la panne : un modèle ne fournissant pas d'information sur l'origine de celle-ci leur est presque inutile car les causes de dysfonctionnement d'un appareil peuvent être innombrables. Un modèle qui fournit des informations sur le composant qui est susceptible de causer une panne permet donc de communiquer avec le métier et d'instaurer la confiance envers l'outil de prédition déployé.

Une autre problématique concerne les contraintes légales et éthiques. Ainsi par exemple, nous pouvons citer un grand opérateur téléphonique qui a mis en place un modèle de Machine Learning pour la détection de fraudes à la souscription de forfaits télépho-

niques. Sa politique commerciale implique que si un dossier est rejeté lors de cette souscription, l'opérateur doit être capable d'en expliquer les raisons au client. Ainsi, le modèle de Machine Learning doit être d'une part interprétable et d'autre part inséré dans un processus qui est défini de telle sorte que le modèle ne peut rejeter un dossier automatiquement. Chaque dossier identifié par le modèle comme douteux est envoyé aux équipes d'analystes fraude pour prendre la décision de rejet manuellement.

II.3.2 – TROUVER LE BON COMPROMIS

Cependant, il faut également être conscient que **favoriser l'interprétabilité d'un modèle d'apprentissage peut altérer sa performance**. En effet, les modèles simples, comme les arbres de décision ou les régressions logistiques ont l'avantage d'être interprétables mais présentent souvent des performances moindres que certains modèles plus complexes comme le gradient boosting. Cette différence de performance est un coût à considérer lors du choix du modèle. Il est également à noter qu'il est possible d'utiliser un modèle initialement peu interprétable dans le but d'obtenir des performances importantes et construire en parallèle des modèles plus simples qui sont eux destinés à l'interprétation des décisions.

En conclusion, on remarque que l'interprétabilité des résultats peut parfois prendre le pas sur les performances d'un modèle. Cependant, elle représente un coût qui doit être estimé si possible dès la phase de POC. Ce coût est à mettre en parallèle de l'objectif principal du modèle, à savoir la performance ou bien la prise de décision. Notons simplement pour finir qu'un compromis peut être d'utiliser en parallèle deux types de modèles : une « boîte noire » (**exemple : XGBoost**) et un modèle plus explicite, le second étant destiné à expliquer en partie les résultats du premier (exemple : entraînement d'un arbre de décision simple sur les résultats du modèle « boîte noire » pour explication des décisions les plus discriminantes).

..... *Nicolas Gibaud*

II.4 | Datavisualisation

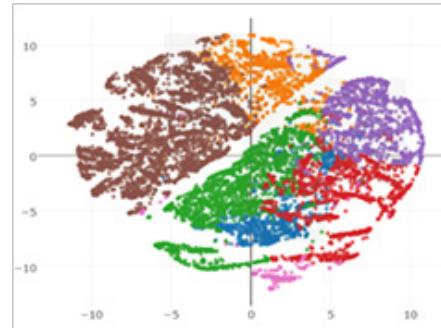
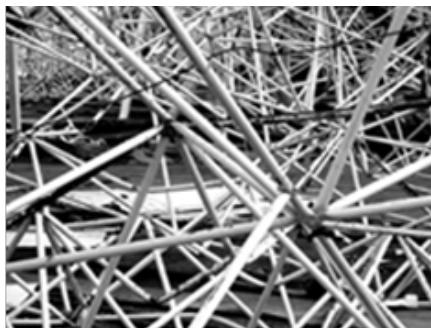
La datavisualisation est l'art de représenter des résultats et des données de façon visuelle. Elle représente un vecteur essentiel de la conduite du changement dans une organisation en facilitant la compréhension de processus complexes par divers interlocuteurs. La data visualisation peut être utilisée **soit comme langage commun, compris de tout le monde, soit comme un outil opérationnel pour les équipes métiers, compris par ceux qui en ont besoin**. Dans ce cas nous parlons de **visualisation « user-centric »**. Ainsi, il est essentiel de réfléchir en amont du processus de mise en production à la pertinence des graphiques et dashboards en fonction des utilisateurs finaux.

La multiplicité des finalités et des interlocuteurs va créer la multiplicité des outils et des technologies. Ainsi, l'écosystème de la datavisualisation est de plus en plus riche mais également de plus en plus complexe. L'enjeu du choix des technologies et de la représentation à utiliser est de taille pour s'assurer de la meilleure efficacité du résultat.

II.4.1 – LEVIER DE LA COMPRÉHENSION

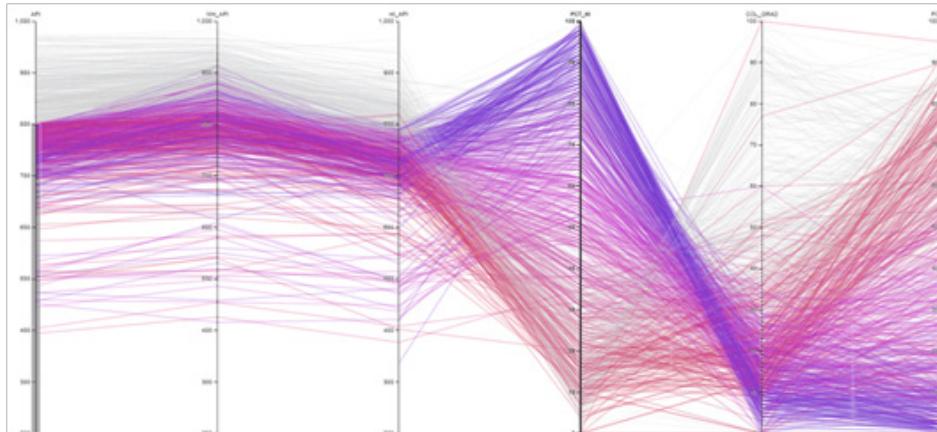
La datavisualisation peut être utilisée dès la phase d'exploration pour comprendre le contenu des données, mais aussi plus en aval pour comprendre et utiliser les modèles. Ainsi le data scientist peut, en un graphique, apporter une information complexe et subtile à un décideur. Métier, IT, analystes, tous les corps de métier impliqués dans un projet ont un langage qui leur est spécifique et qui n'est pas compris des autres entités. Diverses missions dans le monde médical ont montré que les graphiques représentent le meilleur moyen pour faciliter la discussion entre les spécialistes de la donnée et les médecins.

Par ailleurs, l'interaction représente une part essentielle dans la pédagogie. De fait, la construction d'une visualisation interactive donne des possibilités de personnalisation



et de test qui facilitent grandement la compréhension. « Jouer » avec les données est le meilleur moyen de les comprendre et la visualisation offre cette possibilité.

Les grands volumes de données ne sont en rien un frein à la datavisualisation. De nombreuses techniques sont aujourd'hui développées pour offrir la possibilité de visualiser des jeux de données en grande dimension. Ainsi, **les Parallel Coordinates, la courbe d'Andrew, les Self Organizing Maps ou encore le t-SNE offrent différentes possibilités de visualiser des données de grandes dimensions.**



La datavisualisation accompagne aussi le métier dans ses phases de test & learn sur les différents projets mis en place. Elle se présente comme un levier d'amélioration notamment dans la détection d'anomalies ou encore le pilotage de performances.

II.4.2 – LA DATAVISUALISATION « USER-CENTRIC »

Une bonne visualisation est une visualisation adaptée à ses interlocuteurs. La difficulté réside donc dans l'identification des bons utilisateurs pour une visualisation. Une fois que ceux-ci sont définis, l'objectif devient de rendre les graphiques et la visualisation compréhensibles par ces personnes. Dans certains cas, il n'est pas forcément nécessaire de perdre en précision de l'information pour gagner en compréhension globale si les seuls utilisateurs de l'application sont des opérationnels métiers qui maîtrisent déjà la complexité et les enjeux de la donnée. **Il est capital de comprendre à la fois l'utilisateur et son besoin lors de la construction d'une visualisation de données.**

II.4.3 – LE CHOIX DES TECHNOLOGIES

D3js, Tableau, Bokeh, Shiny, Angular.js, Gephi, Leaflet, ... La liste des outils et logiciels de datavisualisation ne fait que grandir, tout comme les possibilités créatives qu'ils offrent. Comment trouver le bon outil, adapté au contexte de production et aux utilisateurs finaux ? Il est possible de distinguer ces outils entre ceux destinés à une utilisation spécifique (**Gephi** pour les graphs, Leaflet pour les visualisations géographiques...) et les outils éditeurs plus génériques comme **Tableau**, **QlickView ou Spotfire**. Viennent s'ajouter à cela des solutions intermédiaires directement implémentées dans les outils du data scientists, au plus près des données, comme **RShiny (R)**, **Plotly (R, Python)** ou encore **Bokeh (Python)**.

Plusieurs aspects peuvent être considérés dans le choix d'une technologie. Ainsi, **la maturité de celle-ci, son coût, sa souplesse (en fonction du besoin), les compétences techniques nécessaires** sont autant de facteurs discriminants pour choisir l'outil le plus adapté. Un critère de choix réside également dans la technicité de l'utilisation de la technologie. Ainsi on distingue les outils destinés aux data scientists (**plotly, bokeh, ...**) des outils développés par les data scientists mais à destination des équipes métiers (RShiny par exemple). En alternative se trouvent les outils déjà développés et utilisés par les équipes métiers (les outils éditeurs cités plus haut). En parallèle, il est important de vérifier les possibilités d'intégration des outils avec les technologies et le matériel choisis dans le processus d'industrialisation. La visualisation doit-elle être interactive ? Les résultats doivent-ils être accessibles sur le Web ? Quelles interactions dois-je prévoir avec mon système d'information ? Beaucoup de questions doivent être posées pour répondre à la question du choix des technologies.

Ainsi, en fonction de la finalité, des interlocuteurs, du processus de production et des compétences techniques internes, le choix sera bien différent mais doit, dans tous les cas, **être réfléchi et mûri en amont dès la phase de POC pour garantir un résultat optimal**. Une datavisualisation adaptée permettra de tirer au maximum la valeur des données. Même si certaines entreprises en ont d'ailleurs fait leur cheval de bataille, il faut garder à l'esprit qu'elle ne reste qu'un outil dont la pertinence est liée à la qualité des données et des modèles sous-jacents.

..... Stéphane Jankowski

II.4.4 – TÉMOIGNAGES

SNCF Digital

Voyez-vous plus la data science comme un outil d'aide à la décision ou comme quelque chose qui à terme viendra automatiser des processus (et peut-être ainsi court-circuiter ce qui est fait par le métier) ?

Aujourd'hui nous aidons plutôt à construire des outils d'aide à la décision. Sur les sujets

de maintenance prédictive et de sécurité, qui constituent jusqu'à présent l'essentiel de notre activité, les modèles probabilistes produits n'ont clairement pas vocation à court-circuiter la prise de décision humaine. Par exemple, nous fournissons des probabilités d'occurrence de pannes de tel type de matériel. Charge au métier de voir comment intégrer cette information dans ses processus et de voir s'il veut automatiser certaines tâches à la lumière de l'information fournie par nos analyses. Nous pouvons éventuellement donner notre avis sur la pertinence ou l'utilité d'automatiser un process, mais nous ne nous substituons pas au métier qui prend la décision.

Safran

Vos modèles sont-ils toujours interprétables par les experts métiers ?

Chez Safran, la data science est à ce jour plutôt considérée comme un outil d'aide à la décision. Nos interlocuteurs sont parfois plus intéressés par la compréhension des phénomènes que par l'aspect prédictif pur, et les modèles interprétables sont alors privilégiés. L'expert a la capacité d'évaluer les résultats et aide ensuite le data scientist à figer les paramètres des modèles.

Il est cependant possible de développer une certaine intuition sur les modèles « boîtes noires ». Ils ne sont donc pas exclus et font partie de notre boîte à outils.

PARTIE III : ARCHITECTURE IT

III.1 | Du POC au pilote

III.1.1 – POC VS PILOTE, QUELLE(S) DIFFÉRENCE(S) ?

Chaque phase d'un projet data science se distingue des autres par son objectif :

- ✖ **POC** : explorer certaines hypothèses (valorisation d'un business case, choix de technologie, identification des features, approche de Machine Learning) en développant un produit informatique. A l'issue du POC, les développements sont jetés et le projet dispose d'un mandat et d'un budget.
- ✖ **Pilote** : valider le fonctionnement dans des conditions réelles d'utilisation avec un périmètre limité en matière de déploiement (sous-ensemble d'utilisateurs, produit ou zone géographique limitée), sur une infrastructure dont le niveau de disponibilité¹ n'est pas assuré. Ces prédictions sont transmises, mais pas nécessairement utilisées par les équipes opérationnelles.
- ✖ **Industrialisation** : généraliser le pilote à l'échelle de l'entreprise en garantissant un niveau de disponibilité maximal, et récolter le ROI. Les prédictions sont utilisées par les équipes opérationnelles et font partie intégrante de leurs fonctions.

Table 1 : Caractéristiques de chaque phase d'un projet data science

	POC	Pilote	Industrialisation
Objectif	Explorer	Valider	Généraliser
Coût	Faible	Moyen	Élevé
Challenge	Conceptuel	Technique	Organisationnel
Utilisateur	Data scientist	Acteurs métier	Acteurs métier
Décideur	Chief Data Officer / Head of Data Lab	Responsable BU	Responsable BU
ROI	Estimé	Vérifié	Obtenu

¹ On définit la disponibilité comme la période pendant laquelle un service est utilisable, par opposition à une période de maintenance.

Par conséquent, les moyens engagés, les difficultés, l'utilisateur final et le commanditaire sont différents pour chacune de ces trois phases (cf. Table 1). Typiquement, le coût d'un POC s'évalue autour d'une dizaine de k€ alors qu'une industrialisation complète peut excéder le million d'euros.

III.1.2 – A QUELLES CONTRAINTES FAUT-IL PENSER LORS DU PASSAGE EN PILOTE ?

Passer en pilote (en vue d'une industrialisation ultérieure) soulève de nouvelles questions par rapport à la phase de POC. Nous pouvons les organiser de manière hiérarchique : d'abord les contraintes imposées par les données (besoins matériels), puis celles de l'utilisateur (besoins fonctionnels), et enfin celles liées à l'entreprise (besoins de sécurité et d'éthique).

Contraintes sur les données

Le stockage

Le premier ensemble de questions est lié au stockage au sens global. Notons d'emblée que passer en pilote (ou en industrialisation) **ne signifie pas forcément utiliser un système de stockage distribué**. Il s'agit d'un choix adapté à la quantité et à l'utilisation des données. Il est crucial à ce stade d'identifier et de **quantifier la limite en capacité de ses ressources matérielles**.

Quel est l'ordre de grandeur du volume total (Mo, Go, To, Po) ? Quel volume maximal serai-je capable de stocker ? Suivant les cas, l'éventail de solutions va du disque dur à l'infrastructure distribuée de type HDFS (ex : Hadoop), RDBMS (ex : Teradata) ou NoSQL (ex : *Apache Cassandra*)².

Quelle est la vitesse d'acquisition ? S'agit-il d'un chargement unique, ou d'un flux de données ? Cette question permet d'estimer le volume total engrangé sur une période de temps donnée, et donc d'anticiper l'espace disque nécessaire à l'enregistrement. Par ailleurs, si le flux de données est vraiment très rapide, on peut se heurter à une limite de vitesse d'écriture sur l'espace de stockage. Dans un tel cas, il faudra faire appel à des technologies spécifiques pour capter ces flux (ex : *Apache Kafka*).

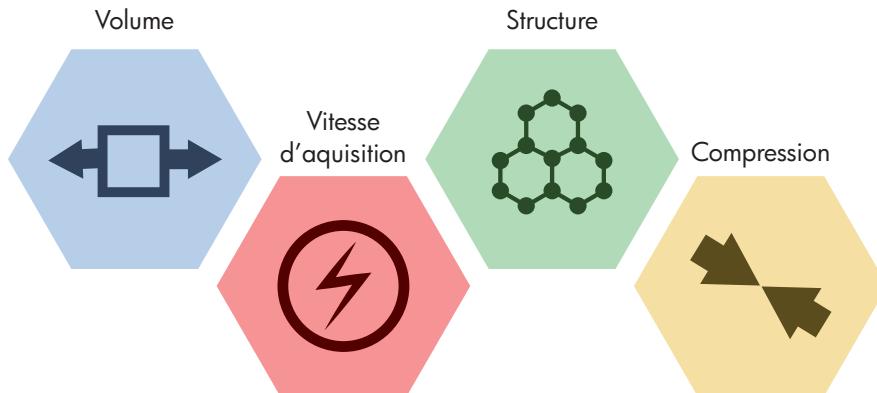
Quelle est la fraction de données structurées (tables) et non structurées (image, son, log) ? Il faut alors choisir un format de stockage le plus adapté aux données, dont on a cité quelques possibilités précédemment.

Est-il nécessaire de tout conserver ? Peut-on agréger la donnée ? La compresser ? Quelle est la fraction de données chaudes (i.e. couramment utilisées) ? Existe-t-il une catégorie de données froides (i.e. peu utilisées) ? Stocker de la donnée peut être onéreux et poser des problèmes de performance pour la recherche³. Une manière simple est donc de choisir de ne pas conserver ou d'agréger la donnée (par exemple dans le temps). Par ail-

² Il faudra alors prendre en compte le facteur de réPLICATION

³ Même si des solutions existent comme elasticSEARCH

leurs, une certaine catégorie des données doit parfois être stockée (données médicales), mais en pratique est très rarement utilisée (mise en archive). Il devient alors intéressant de considérer certaines données comme mortes, et d'autres actives, et de **choisir des stratégies de stockage différenciées**.



Le traitement des données

La nature des données est une information nécessaire, mais pas suffisante pour déterminer l'infrastructure technique de leur gestion. **Il est nécessaire d'identifier quels types d'opérations on souhaite appliquer.**

Les opérations sont-elles globales ou locales ? En d'autres termes, les opérations s'appliquent-elles sur une partie restreinte des données (ex: insertion d'une nouvelle donnée), ou nécessitent la connaissance de toute l'information en même temps (ex: calcul de la moyenne).

Quelle est la fréquence de lecture/écriture ? Si la fréquence est faible (par exemple une fois par jour), on peut considérer stocker la donnée sur des espaces dédiés à bas coût financier, mais dont l'accès est beaucoup plus lent. Au contraire, si la fréquence est élevée, il vaudra mieux choisir un système de stockage «chaud» comme les bases de données.

La scalabilité

Enfin, il est nécessaire de se demander si les contraintes précédemment décrites sont susceptibles d'évoluer dans le temps. Est-ce que le volume des données va drastiquement grandir (de plusieurs ordres de grandeur) ? Cette information peut avoir des conséquences profondes sur les choix d'architecture et va imposer de prêter attention au temps de calcul (au-delà de l'espace disque de stockage). En effet, le temps de calcul grandit généralement avec la taille des données. **On parle alors de complexité algorithmique.** Par exemple, un temps de calcul qui augmente linéairement avec la taille n des données, a une complexité en $O(n)$. Quel que soit le temps de calcul pour des faibles

tailles, la complexité algorithmique finit par influencer très largement le temps de calcul à haut volume. Notons que nous parlons ici de la scalabilité d'un projet seul, et qu'il faut donc ajuster dans le cas où des ressources sont mutualisées avec d'autres projets.

Autant que possible, il faut estimer la complexité algorithmique afin d'anticiper les besoins en ressources, et faire les opérations nécessaires pour assurer **une borne supérieure au temps de calcul**⁴. La même logique peut s'appliquer au chargement en RAM, c'est à dire qu'il est recommandé d'anticiper **la mémoire vive nécessaire au calcul** lorsque le volume des données croît. Il n'est pas rare de prévoir un dépassement des exigences en temps et en RAM lorsque le volume de données maximal est atteint. Anticiper ce fait, c'est pouvoir changer d'algorithme, et parfois aussi changer la stratégie de stockage des données en amont. Quelles solutions sont alors possibles ?

Tout d'abord, on peut appliquer une **scalabilité verticale**, c'est-à-dire augmenter le nombre de processeurs et/ou la mémoire RAM. Par opposition, on peut appliquer une **scalabilité horizontale**, c'est-à-dire mettre en parallèle plusieurs machines pour répartir la charge de calcul. Cependant, ces éléments supposent que le calcul est distribué et découpable en sous-ensembles cohérents (comme pour du map reduce).

En complément, on peut pré-calculer certaines quantités pour accélérer le temps de traitement. Dans le cas où il n'est pas nécessaire de réentraîner le modèle, la solution d'utiliser un modèle pré-entraîné est viable, et permet d'obtenir des performances modèles connues à l'avance. Bien entendu, toute la difficulté de cette démarche réside dans la recherche d'un modèle pré-entraîné qui répond au besoin.

Contraintes expérience utilisateur

En POC, seules les équipes de développement et d'analyse manipulent le service. En pilote, la situation est différente puisque ce sont des acteurs métiers qui interagissent avec ce dernier hors du laboratoire. Généralement, l'équipe de développement **se munît alors d'une série de tests qui reproduit toutes les erreurs possibles d'utilisation** et contrôle le comportement du service. De la même manière, **il faut aussi garantir la disponibilité du service** (aussi connue sous le nom de **uptime**). Cela peut potentiellement se traduire par la présence d'une **équipe de maintenance** capable d'agir en cas de perte du service. Au-delà de ces points obligatoires, on peut également vouloir améliorer le confort de l'utilisateur en garantissant un délai de réponse maximal, ou autrement dit, un temps de calcul maximal, souvent de l'ordre de la seconde pour un utilisateur humain. De manière générale, on parle d'améliorer l'expérience utilisateur (UX design).

Par ailleurs, un passage en pilote implique généralement plusieurs utilisateurs interagissant en même temps sur des données. On parle alors de **multi-threading**. Cela peut avoir des impacts de performance sur le réseau, le temps de calcul, mais également sur la cohérence des données.

..... *Long Do Cao*

⁴ Cette estimation est idéalement déjà effectuée en phase de POC

III.2 | Les patterns d'architecture

III.2.1 – LES TYPOLOGIES DE BESOINS

Qui dit industrialisation d'une application dit nécessité de l'intégrer dans le cadre d'une architecture adaptée. Une architecture adaptée doit être capable de supporter les contraintes liées à l'exploitation opérationnelle de l'application tout en étant **la plus légère et souple possible**. La légèreté et la souplesse de l'architecture sont en effet des dimensions cruciales car elles aideront à minimiser la dette technique et ainsi faciliter l'exploitation de l'application, sa maintenance et ses potentiels demandes d'évolution (scalabilité, refonte / migration technologique...).

Pour supporter des sujets autour de l'exploitation de données les technologies « Big Data » comme Hadoop, Spark, les technologies *NoSQL* ou encore *Kafka* connaissent depuis quelques années un franc succès dans l'entreprise. **Car, en plus d'être open source, elles sont suffisamment généralistes pour porter 99 % des cas d'usage et ce, entièrement de manière scalable et fault-tolerant.** Sur le papier, il suffit donc avec ces technologies d'ajouter de nouvelles machines pour constater une réduction du temps de calcul de leurs requêtes, une augmentation de sa capacité de stockage et une meilleure gestion de large flux de données temps-réel de manière linéaire. Ainsi les architectures principalement axées autour de ces technologies Big Data (comme par exemple l'architecture *SMACK* : *Spark*, *Mesos*, *Akka*, *Cassandra*, *Kafka*) sont très performantes et relativement simples à concevoir mais appliquer et déployer ce pattern à toute application peut s'avérer peu efficace...

En effet, bien que ces technologies règlent la question centrale de la scalabilité, elles viennent avec un prix qu'il est nécessaire de prendre en compte, le coût de leur mise en place et de leur exploitation. Ces technologies restent relativement jeunes et demandent aux développeurs d'avoir un niveau de compréhension très fin de leurs fonctionnements internes. **Se doter de ces technologies sans les bonnes ressources peut alors se traduire par un constat d'une baisse sensible de productivité.**

Il est ainsi nécessaire, lors de la conception d'une architecture, d'arbitrer constamment entre le réel besoin de scalabilité et l'utilisation de briques plus matures ou simples. Pour identifier une architecture adaptée à son cas d'usage, une approche possible est d'utiliser un pattern de référence dans la mise en production d'application Big Data comme le **pattern d'architecture Lambda**. Ce pattern très générique porte souvent plus de briques fonctionnelles que nécessaire. On peut donc s'en servir comme base et l'épurer en fonction des réels besoins de l'application.

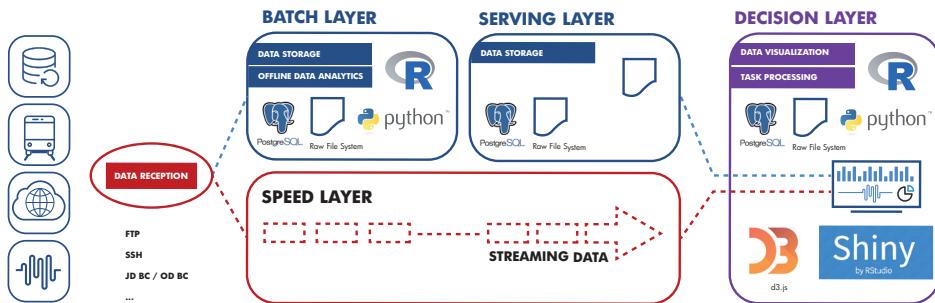


Figure 1 : Pattern d'architecture Lambda

L'intérêt du pattern d'architecture Lambda est de présenter de manière découpée les différentes briques/technologies de gestions des données selon la « température » de traitement associée. Cette température peut être :

- ✖ « **Froide** » : si la gestion des données peut se faire en mode batch - historisation de la donnée, traitement analytique à fréquence quotidienne ou plus importante...
- ✖ « **Chaud** » : si la gestion des données doit se faire en temps réel (ou en micro-batch) : gestion de flux chauds (logs, objets connectés...), traitements/indexation en temps réel...

Les briques de l'architecture Lambda peuvent communiquer entre elles et peuvent être activées ou désactivées en fonction des cas d'application. Il est alors possible d'appliquer un ensemble de technologies à l'état de l'art sur ce pattern. L'ensemble technologique cible pourra alors être du plus complet avec des technologies scalables à de larges volumes de données (exemple : architecture hybride **Hadoop + SGBDR**. Cf Figure 2) au plus simple pour la gestion de volumes de données plus modestes (exemple : simple serveur de calcul avec outils de manipulation de données - Python, R... cf. : Figure 3).

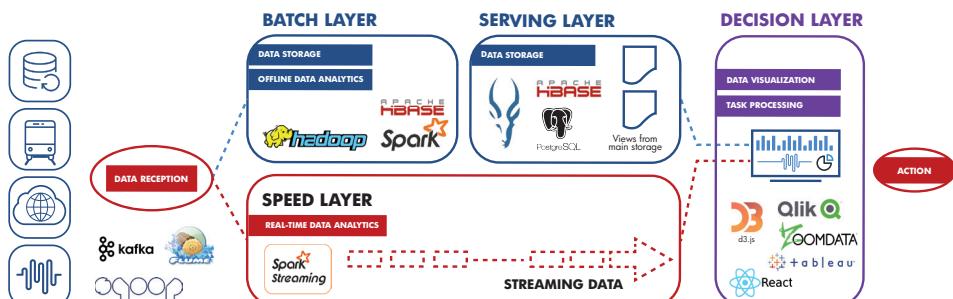


Figure 2 : Lambda architecture complète sur Stack Hadoop + SGBDR / NoSQL

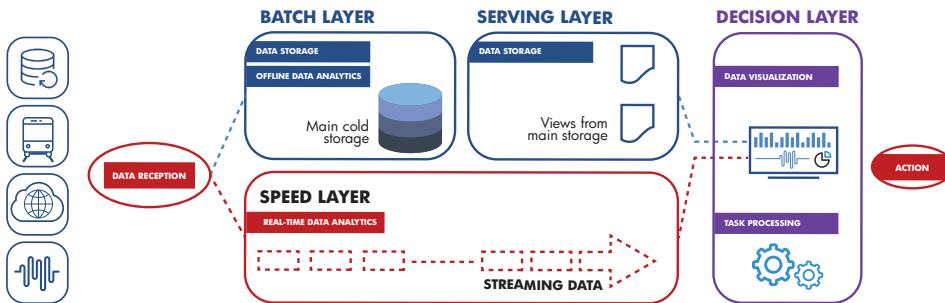


Figure 3 : Architecture simple (sans speed layer) pouvant être portée par un seul serveur

Ces deux illustrations représentent deux cas d'implémentations « extrêmes » du pattern Lambda :

D'un côté en *Figure 2*, l'architecture est entièrement scalable à des **volumétries massives** (largement supérieures au Téraoctet) et permet de traiter une quasi exhaustivité de cas d'application mais celle-ci peut être lourde à mettre en place et à maintenir. En effet, la réaliser et l'exploiter nécessite au minimum une équipe de **2 à 3 data engineers expérimentés** entièrement dédiée à la gestion des flux de données en complément d'une équipe de **2 ou 3 data scientists** dédiée à l'exploitation des données en traitement distribuées.

D'un autre côté en *Figure 3*, l'architecture est quasiment restreinte à de l'exploration et ou à du traitement par batch, le tout **difficilement scalable au-dessus du Téraoctet**, mais elle présente l'avantage de pouvoir être rapidement mise en place et facilement exploitable. Un unique **data scientist aguerri** pourrait être autonome pour exploiter les données et produire des applications analytiques de manière agile sur cette architecture. Ainsi, avec ces deux exemples extrêmes à l'esprit au moment de choisir une architecture pour une application analytique (idéalement déjà éprouvée par une phase POC), il est possible d'identifier une architecture adaptée en trouvant un équilibre entre agilité d'exploitation et capacité de montée en charge de l'architecture.

Qui plus est, de manière générale, **il sera souvent plus facile de « complexifier » une architecture simple que d'alléger une architecture surdimensionnée**. Ainsi, on ira vers des composants Big Data seulement dans le cas où les réponses aux questions types suivantes sont affirmatives.

Dans notre application, avons-nous **vraiment** besoin :

- ✗ D'ingérer un flux de données de l'ordre de **plusieurs centaines d'évenements par seconde (ou plus)** ?

- Si oui, alors une application comme **Apache Kafka** peut par exemple être considérée.
- Si non, alors considérer ingérer les données en batch, ou par INSERT concurrents d'une base relationnelle type « **OLTP** ».

✖ D'un espace de stockage historique distribué et **scalable** ?

- Si oui, alors des technologies comme **Hadoop HDFS, Amazon S3 et des bases NoSQL** peuvent être considérées.
- Si non, considérer éventuellement un **SGBDR « simple » voire le File System** d'un serveur.

✖ De distribuer les traitements des données **sur plusieurs nœuds** ?

- Si oui, alors on peut envisager utiliser des technologies comme Hadoop MapReduce ou Apache Spark
- Si non, considérer par exemple des traitements **SQL** et/ou analytiques avec **R ou Python**.

✖ D'exposer des données **non supportées par un SGBDR classique** ? (i.e. non représentables en table et/ou organisées sous forme de clés valeurs)

- Si oui, alors considérer des technologies **NoSQL** comme par exemple **HBase, Cassandra, ou Elastic Search** en fonction du type de données,

✖ De traiter les données **en temps réel** ?

- Si oui, alors considérer des technologies comme **Spark Streaming** voire **Apache Storm**.

III.2.2 – CLOUD

A moins que le POC ait un tel retour sur investissement que son financement soit en mesure de porter un projet d'architecture complet, **nous déconseillons de commencer par établir une infrastructure Big Data en interne**. La complexité des couches logicielles (Hadoop, Spark), la technicité des développements et les coûts d'infrastructures sont souvent élevés et sous-estimés. Par ailleurs, les compétences particulières d'administration de systèmes distribués font que l'exploitation d'une telle architecture est souvent délicate à monter et requiert un niveau d'expertise qui n'est pas à la portée de tous.

La définition de l'architecture est une entreprise de longue haleine qui doit être mutualisée. Le temps de déploiement d'une telle architecture incluant la formation des exploi-

tants peut dans certains contextes dépasser la demi-année. Hors, les résultats du POC ont rendu l'entreprise impatiente de déployer le système et il est difficile de piloter avec sérénité un tel projet avec une direction exigeant des résultats rapides.

Heureusement, il existe une solution alternative trop souvent négligée et pourtant très utile : le passage par le cloud computing. Les hautes performances et les possibilités de stockage illimitées qu'offre le cloud en font un outil particulièrement intéressant pour les usages du Big Data. Il facilite la scalabilité et sa flexibilité permet d'expérimenter, de revenir en arrière, et de déployer les premiers projets en concentrant les efforts sur la data science elle-même.

Il existe en réalité trois obstacles (et ce sont les seuls) au passage sur le cloud :

- 1. Des contraintes réglementaires** (une banque ne peut pas mettre certaines données sur le cloud).
- 2. Des contraintes fortes de disponibilité** qui ne sont pas couvertes par un contrat. En général, la disponibilité du cloud est meilleure que celle d'une infrastructure locale MAIS un opérateur cloud peut se réserver le droit de stopper des composants quand il le veut. (Si votre application Big Data sert à transporter des personnes lors des Jeux olympiques, vous ne pouvez tout simplement pas prendre ce risque).
- 3. Une sensibilité aux ruptures réseau** (pas de réseau = pas d'application). Habitué à la 4G et au réseau haut débit, on considère les liaisons réseau comme allant de soi. C'est une erreur.

Les offres cloud se subdivisent en trois niveaux :

- ✖ **IaaS (Infrastructure as a Service)** : le fournisseur loue une infrastructure informatique sur laquelle le client va gérer toute la couche applicative.
- ✖ **PaaS (Platform as a Service)** : en plus de l'infrastructure, le fournisseur se charge de maintenir l'environnement d'exécution des applications. Aujourd'hui, la plupart des plateformes de traitement distribué sont disponibles (*Hadoop, HDFS, Yarn, Spark...*). (Ex : *AWS EMR, Azure HDInsight, Google Cloud Platform, etc.*)
- ✖ **SaaS (Software as a Service)** : c'est le plus haut niveau, dans lequel un éditeur de logiciel propose des solutions et des services en plus d'*Hadoop* pour les clients qui souhaitent se mettre à la data science mais n'ayant pas forcément les compétences adéquates (ex : *Qubole Data Service, IBM SPSS Modeler Gold, etc.*). Ces services sont généralement eux-mêmes hébergés sur le IaaS.

..... *Abdellah Kaid-Gherbi / Matthieu Vautrot*

III.3 | Edition de logiciel, comment choisir ?

III. 3.1 – A QUEL MOMENT CHOISIR

Si les POCs peuvent se faire sur une architecture logicielle 100 % open source déployée sur un serveur puissant ou une infrastructure cloud IAAS, vient un moment où l'on doit se pencher sur la pérennité du code et l'élaboration d'un socle technique commun. Un des premiers pièges est de vouloir traiter ce sujet immédiatement. Or, aux prémisses des initiatives Big Data, les équipes ont besoin d'explorer des choix, de formuler des hypothèses, de revenir en arrière et de se faire une idée sur l'utilité de ces nouveaux modèles. **Ce n'est qu'après avoir passé ce cap d'expérimentation que vous êtes en mesure d'exprimer un ensemble de besoins cohérents pour les reporter dans un cahier des charges.**

III.3.2 – QUI CHOISIR

Comparativement aux précédentes vagues de l'informatique, le paysage des fournisseurs s'est fragmenté et les éditeurs classiques ont perdu du terrain face à des nouveaux entrants.

Deux phénomènes sont à l'origine de cette transformation du marché de l'édition logicielle : Le développement de l'open source poussé par les géants du digital et l'apparition d'opérateurs de plateforme cloud. Les géants du digital contribuent massivement au développement de l'open source et mutualisent leurs efforts pour fiabiliser les composants élémentaires de leur infrastructure. Leur modèle économique étant différent de celui d'un fournisseur de solution technologique, ils ont intérêt à ce que leur couche logicielle soit la plus utilisée et à associer les meilleurs développeurs à leurs fondations open source. Ils mettent donc gratuitement à disposition des autres développeurs leurs composants. Il en résulte que les efforts mutualisés d'un Facebook, Yahoo ou IBM excèdent très largement les capacités de R&D d'un éditeur. Dans certains domaines et en particulier pour le Big Data, **l'open source est devenu tout simplement largement meilleur qu'une solution commerciale.**

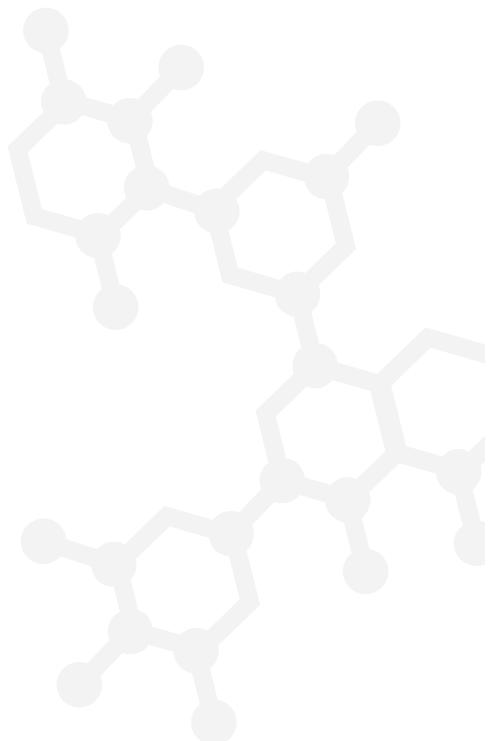
Cette irruption de l'open source a fait émerger une nouvelle catégorie d'éditeurs du logiciel comme *Cloudera, MapR et HortonWorks*. Ces sociétés réalisent **un assemblage cohérent de technologies open source (les distributions)**, orientent et consolident les travaux de la communauté et assurent le déploiement et l'expertise technique pour les clients de ces distributions. Pour des entreprises qui n'ont pas la capacité de dédier quelques experts techniques à la compréhension de ces couches, ce choix est tout à fait indiqué. Bénéficier d'un ingénieur expert sur une technologie de pointe vous fera gagner un temps précieux lors de vos projets. L'avantage de passer par Cloudera ou Hortonworks, c'est que l'expert technique a contribué le plus souvent au code open



source incorporé dans la distribution. Au vu de la complexité technologique d'un framework comme Hadoop, cette expertise est irremplaçable. Les éditeurs classiques ont une stratégie un peu analogue en assemblant composants open source et propriétaires. D'autres acteurs ont aussi émergé et concurrencent sérieusement les éditeurs habituels. **Il s'agit des fournisseurs de plateforme PAAS et IAAS.** Dans bien des cas, une solution PAAS/IAAS est envisageable, notamment pour le passage en industrialisation en raison de la souplesse et la scalabilité offerte par le cloud.

Enfin, les derniers apparus sont **les fournisseurs de services SAAS**. De nombreuses start-up proposent aujourd'hui des plateformes technologiques dédiées à un ou plusieurs

cas d'usage. On trouve ainsi des spécialistes de la détection de fraude, des fournisseurs de publicité ciblée, ou encore des entreprises comme Palantir, qui déploient des verticaux dédiés à un secteur d'activité donné. Cette alternative SAAS a l'avantage de fournir a priori une solution si ce n'est clé en main, en tout cas déjà proche du cas d'emploi. Sur ce type de marché, les fournisseurs de logiciels ont une carte à jouer.



..... *Emmanuel Manceau*

III.3.3 – QUELS CRITÈRES

Voici un petit tableau de synthèse qui décrit les choix possibles et les alternatives.

Type de solution technique	100 % Open Source	Distribution Open Source (Horton Works, Cloudera, MapR)	Solution logicielle classique (SAS, SAP, Oracle, Qlick, Tableau)	IAAS (Amazon EC2, OVH, bluemix Watson...)	AAS (Ge Predix, Windows IBM, Bluemix Watson...)	Avantages du PaaS +
Liberté, couverture des besoins	Mêmes avantages que l'open source Accès à des ressources expertes. Capacité à accompagner les projets Accès à un support	Mêmes avantages que l'open source Accès à des ressources expertes. Capacité à accompagner les projets Accès à un support	Richesse des composants graphiques Les DSJ savent travailler avec Stabilité Accès à un support	Mêmes avantages que l'open source + Performance Stabilité Investissement initial limité	Mêmes avantages que les éditeurs et IAAS + Performance Investissement initial limité	Faible coût de build
Faiblesses sur les composants graphiques Niveau technique requis Instabilité, pérennité			Faible degré d'innovation Dépendance Risque d'être dépassé Tentation de forcer les besoins pour qu'ils conviennent au produit	Coût de long terme de l'engagement disponibilité délictueuse à contractualiser Dépendance fournisseur et au réseau Internet	Couverture des besoins + Mêmes inconvenients que les éditeurs +	

III.4 | Data lake

Un data lake est un référentiel de stockage de données historiquement en lien étroit avec la technologie Apache Hadoop et son écosystème. Concept incontournable des architectures Big Data, le data lake bouscule les conventions et usages des systèmes de bases de données et effectue de nombreuses promesses.

III.4.1 – POURQUOI UN DATA LAKE ?

L'enjeu d'un data lake est de créer de la synergie entre toutes les données de l'entreprise et de rendre possible l'**exploration et l'analyse de toutes les dimensions business**. : Le data lake est le lieu idéal pour croiser les données, habituellement séparées les unes des autres dans des silos organisationnels et techniques, et identifier des nouveaux usages (le marketing s'empare des données des véhicules connectés, les ingénieurs peuvent enfin croiser les informations industrielles avec celles des usages clients). La mise en relation de données hétérogènes sans préjuger d'un usage a priori est néanmoins contraignante techniquement et peu de solutions techniques permettaient de le faire avec un coût et une complexité raisonnable.

Pour arriver à ces résultats les solutions actuelles (*SGBDR, Data Warehouse...*) ne permettent pas de relever les nouveaux challenges :

- ✖ **Environ 80 %⁵ des données produites sont semi-structurées ou non structurées**, et ne pourront pas être intégrées dans une solution classique. Aujourd'hui, toutes les sociétés interagissent avec leurs clients et fournisseurs sur Internet et les réseaux sociaux et ont intérêt à maîtriser et stocker les informations qui les concernent. Ces nouvelles typologies de données (vidéo, image, texte...) poussent à un nouveau mode de stockage.
- ✖ De nombreux cas d'usage **nécessitent une forte réactivité** : un technicien devra obtenir sans attendre le contexte et les réparations à opérer sur une machine, bloquer une transaction suite à une détection automatique de fraude. Une architecture basée sur Hadoop permet cette réactivité à un tarif abordable, en parallélisant les traitements sur un cluster de machines à bas coût.
- ✖ L'explosion du volume de données pour une entreprise peut être prohibitif sachant que le coût d'acquisition et de maintenance d'1 To de données est de 35 000 \$ dans un Data Warehouse traditionnel. **Ce coût tombe entre 1 500 \$ et 3 500 \$⁶ sur environnement Hadoop.**

⁵ «File Systems Define the Future for Managing Storage», Noemi Greyzorf and Richard L. Villars, December 2008, IDC #215463

⁶ <https://www.mapr.com/gigaom-hadoop-data-warehouse-interoperability-report>

Cependant, pour des données structurées et non changeantes, les systèmes de stockage classiques comme les Data Warehouse restent la référence en termes d'efficacité. Nous conseillons la cohabitation des deux univers entre stockage structuré traditionnel d'un côté et données massives et non structurées de l'autre.

III.4.2 – DE LA BASE DE DONNÉES RELATIONNELLE AU DATA LAKE

L'approche historique de la base de données relationnelle

La base de données relationnelle permet de stocker des données structurées dans un système lui-même structuré et hiérarchique du fait de la définition d'un modèle de données relationnel. Elle se caractérise par une bonne visibilité des liens existants entre les données et par une rigueur dans la façon d'organiser le stockage des données.

Cependant, les avantages d'une approche par base de données relationnelle sont également ses faiblesses. **Les SGBDR pêchent par manque de souplesse et d'adaptabilité** aux use cases décrits précédemment. Quand la donnée évolue et que sa structure change, il est lourd et complexe de faire évoluer le modèle : la structure dépend très fortement du cadre défini en amont à la création, on parle de l'approche « schema on write ».

L'arrivée du Data Warehouse

Le principe de Data Warehouse a fait son apparition dans les années 90. L'idée est de rassembler toutes les données relationnelles dans un même entrepôt dans le but de désolidier la donnée. La structure en étoile souvent utilisée permet de garder l'information la plus précise au cœur du Data Warehouse et de la compléter par d'autres informations qui gravitent autour.

L'IT s'affranchit en partie de la forte structuration de la base mais est toujours **dépendant de la structuration des données**. Nous restons dans le paradigme du « schema on write » ..

L'arrivée du concept Data lake

Le data lake représente aujourd'hui la meilleure alternative pour le stockage de données structurées et non structurées. Après l'apparition de Hadoop et des systèmes de fichiers distribués, stocker des données changeantes et dont on ne connaît pas l'utilisation a priori est devenu possible. **On passe d'une structure hiérarchique (verticale), à une structure plate (comme un lac !).**

Le data lake vient se substituer au trio ETL - Data Warehouse - Datamart qui paraissait intouchable quelques années auparavant. Il crée un environnement où la **structure est changeante et s'adapte aux besoins de l'analyse** (on parle de « schema on read » en opposition au « schema on write » des bases de données relationnelles). Ainsi,

la structure du data lake facilite l'utilisation, l'analyse et la valorisation de la donnée de l'entreprise.

III.4.3 – RÉUSSIR LA MISE EN PLACE DE SON DATA LAKE

Nous conseillons systématiquement d'avoir une approche de constitution et d'intégration d'un data lake par méthode itérative. Généralement **la création d'un data lake doit se faire non pas en remplacement d'un SGBDR mais en complément**. Une fois constitué et lié au SGBDR, il pourra alors en fonction de son stade d'évolution :

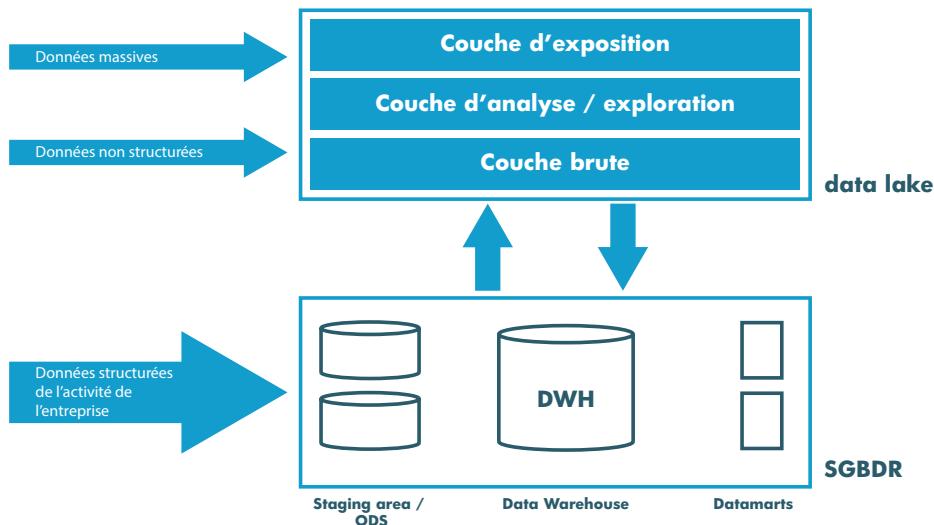
- ✖ permettre l'accès à des cas d'usage non portés par la base traditionnelle (analyses de données massives et / ou non structurées, calculs et analyses lourds) ;
- ✖ soulager les fonctionnalités existantes du Data Warehouse (offloading des traitements ETL, historisation des données non nécessaires dans le DWH).

La constitution d'un data Lake doit aussi s'inscrire dans le cadre plus général de la gouvernance des données, sujet stratégique dont la mission est de :

- ✖ mesurer et améliorer la qualité du patrimoine informationnel (détecter des signaux faibles avec des données fausses est impossible) ;
- ✖ créer du lien entre les groupes de données, afin d'éviter de retomber dans une architecture silotée ;
- ✖ sécuriser le data lake, notamment en gérant les accès par utilisateur ;
- ✖ vérifier les aspects juridiques de désensibilisation et d'anonymisation.

Enfin, pour d'éviter l'écueil du Data Swamp (« déversoir » dans lequel le risque de perdre des données est fort), il est important d'avoir une hygiène irréprochable dans l'organisation du cycle de vie des données sur data lake. Il est alors important de se doter d'une organisation capable de séparer les espaces logiques de données suivants :

- ✖ **couche de données brutes** : pour la réception des flux de données en l'état sans modification ;
- ✖ **couche d'exploration** : pour l'exploitation des données brutes et potentiellement la stabilisation de versions de données plus propres ;
- ✖ **couche d'exposition** : pour figer des vues de données prêtes à être exploitées par des applications métier.



Intégration d'un data lake en complément d'un SGBDR

..... Stéphane Jankowski / Youssef Benchekroun

III.4.4 – TÉMOIGNAGES

Bouygues Telecom

Sur quelle architecture informatique repose votre démarche pour l'instant ?

Nous disposons de quatre clusters Hadoop, dont un à vocation exploratoire qui centralise les données hétérogènes. Mais nous arrivons aux limites de cette approche. Afin de tendre vers une démarche plus industrielle, nous étudions l'opportunité de nous doter d'un lac de données où seront rassemblées davantage de sources différentes. Cette initiative a été amorcée en fin d'année dernière. L'objectif est d'assurer la pérennité de ce qui a été bâti depuis trois ans et d'ajouter de nouvelles données pour améliorer nos analyses. Pour le moment, nous n'avons pas encore tranché entre choix d'une solution cloud ou d'une infrastructure propre.

Criteo

Sur quelle architecture repose votre environnement Big Data?

Nous sommes équipés de deux clusters Hadoop de 1000 nœuds chacun, qui comptent parmi les plus grands d'Europe où siègent nos activités décisionnelles et la construction des modèles prédictifs. Il s'agit de nos propres serveurs, une solution cloud étant difficilement envisageable du fait de la nécessité de sauvegarder 50 pétaoctets de données.



Outre Hadoop, nous utilisons les technologies traditionnelles telles que Kafka, Storm, Spark. En frontal, toute une série de briques Web génère des flux d'information dirigés vers un entrepôt de données pour y être traités de façon centralisée. Elles sont réparties

selon des rôles bien précis : le lien avec les plateformes d'enchères, la réception des catalogues produits de nos clients, la génération d'images, la pose de cookies... Pour servir nos clients partout sur la planète, nous disposons de sept centres de données capables de garantir les temps de réponse exigés.

A quelles contraintes êtes-vous soumis en matière de temps de réponse aux enchères?

Les temps de réponse sont encadrés par des contrats : si l'on répond en plus de 10 ms, le délai est expiré. Si la situation se reproduit trop souvent, nous ne sommes plus sollici-

tés. Il faut à tout prix éviter cette situation de « black-listé ». Si l'on soustrait le temps de connexion, la latence réseau, cela nous laisse donc un délai très court, entre 10 et 30 ms environ, pour appliquer nos modèles de prédiction. Une fois l'enchère gagnée, il faut faire en sorte que la bannière s'affiche rapidement afin que l'expérience de l'internaute reste agréable. Nous disposons d'un laps de temps d'environ 15 à 20 ms pour choisir le produit à présenter et quelques millisecondes pour générer la bannière.

Menez-vous des travaux de recherche d'optimisation du temps de réponse ?

Oui, car se retrouver « black-listé » pour une heure d'enchères signifie une perte sèche de chiffre d'affaires. C'est donc pour nous un événement grave. Mais cela s'apparente plutôt à de l'ingénierie classique qu'à de la recherche académique. Appliquer un modèle prédictif est quasiment déterministe : nous savons combien de temps cela prendra. Dans le domaine de la gestion opérationnelle d'un système distribué, il s'agit plutôt d'un travail au long cours. C'est toute une culture opérationnelle à acquérir de rationalisation de la plateforme qui est de plus en plus présente au sein de nos équipes.

Safran

Avez-vous pensé l'architecture en amont (dès le POC) ? Et a-t-elle changé depuis ?

Pour réaliser nos différents travaux, nous avons mis en place différentes plateformes. La première nous sert de banc d'essai pour les différentes technologies. La seconde nous permet de réaliser les POCs sur des données froides. La troisième est une plateforme de pré-industrialisation pour les projets pilotes sur des données chaudes. Enfin la plateforme de services nous sert de portail pour les clients afin qu'ils prennent en main les outils et appréhendent les use cases.

A ce stade, nous préférons utiliser les outils open source classiques de la data science (R, Python, Scala, SQL, Spark...) que nous assemblons en fonction des besoins.

Il est donc compliqué de choisir une solution commerciale *a priori*. Qui plus est, nous souhaitons monter en compétences et maîtriser les composants avant d'acheter une solution propriétaire.

PARTIE IV : Vie quotidienne d'un modèle

IV.1 | Cycle de vie

IV.1.1 – UN MODÈLE, VISION MATHÉMATIQUE APPROCHÉE DE LA RÉALITÉ

Un modèle est un outil de transformation des observations en quantités mathématiques, physiques ou numériques, d'un état instantané vers une histoire. Le but d'un modèle est de capturer, à partir d'observations passées, les éléments importants permettant de prédirer le comportement futur de l'objet modélisé. **Un modèle n'est pas conséquent jamais parfait, ni totalement représentatif de la réalité.** Lorsque l'on utilise les modèles comme un outil pour décrire le réel (des quantités mathématiques vers des observations), on parle de modèle prédictif. Un modèle est alors jugé performant ou pertinent lorsque la prédiction est proche de la réalité.

Dans le cadre d'un projet de data science, les observations sont incarnées par les données collectées, les quantités mathématiques par les variables explicatives. Les équations mathématiques sont quant à elles traduites en algorithmes informatiques. Le processus de modélisation, c'est-à-dire la création d'un modèle, implique généralement 4 phases :

Construction du dataset : extraction et nettoyage des données. Etant donné la large variété des données existantes, cette étape peut représenter une grande fraction d'un projet data science (scrapping, processus ETL, inférence des valeurs manquantes).

Feature engineering : identification et/ou création des variables explicatives. Ces dernières vont capturer l'essence de l'objet à modéliser.

Algorithmes : choix de l'algorithme informatique pour combiner les variables explicatives et faire une prédiction. A ce stade, plusieurs algorithmes sont généralement confrontés. Le résultat de cette phase est généralement un ensemble de paramètres de l'algorithme (ex : les coefficients d'une régression linéaire). D'autre part, chaque algorithme possède souvent des hyper-paramètres non contraints par les données (ex : profondeur d'un arbre dans une random forest), mais qui doivent être réglés selon une méthodologie précise pour optimiser les résultats et la vitesse d'exécution.

Evaluation : mesure de performance. C'est à ce stade que l'on définit comment juger de la qualité d'un modèle. En d'autres termes, on quantifie la proximité de la prédiction avec la valeur réelle (écart entre la prévision et la réalité). Une métrique pour un modèle ne va pas refléter toutes ses qualités ; il faut veiller à prendre en compte le contexte dans lequel le modèle s'exécute.

En termes de reproductibilité des résultats, les trois premières phases sont en fait indissociables : changer l'une des briques peut changer complètement le résultat des algorithmes. De ce constat, nous qualifierons donc de modèle l'ensemble de ces trois étapes.

IV.1.2 – LA VIE D'UN MODÈLE « PRIMITIF »

Dans un processus d'industrialisation, **le modèle « vit »** : il évolue avec le temps et suit un cycle de vie complexe.

La naissance est actée lorsque le modèle est branché aux systèmes informatiques et aux processus de l'organisation. A ce stade, les éléments suivants sont figés : le code source du modèle, les hyper-paramètres de l'algorithme, les paramètres de l'algorithme, et donc a fortiori les données d'entraînement.

On peut matérialiser « l'âge » d'un modèle en lui attribuant un **numéro de version** (ex : v1.0), pratique couramment utilisée dans le monde du développement. A chaque changement du numéro de version, on sauvegarde une copie complète, à cet instant, du modèle.

La moindre modification d'une des composantes d'un modèle constitue une évolution, aussi petite soit-elle. Ainsi, même si l'on change seulement la valeur d'un paramètre, le numéro de version doit manifester ce changement, afin d'assurer la reproductibilité du processus. Notons qu'un changement dans les données va aussi impacter l'état interne du modèle.

Tant que les modifications touchent à l'état interne ou au code interne, on considère que le modèle évolue. Autrement dit, il continue de vivre. Cependant, changer une brique de code externe signifie que l'intégration du modèle à son environnement est compromise. Cela se traduit alors par une interruption de service et une refonte du code interne. **Le modèle est en ascension.**

IV.1.3 – TÉMOIGNAGES

Bouygues Telecom

Comment garantissez-vous la pérennité des modèles, une fois ces derniers en production ?

J'ai le sentiment que tous les projets ne demandent pas la même implication de la part de nos équipes : certains modèles requièrent une simple maintenance des données et systèmes, d'autres nécessitent d'être remodelés régulièrement par les « data scientists ». En tout cas, je tiens à ce que l'on constitue un patrimoine d'algorithmes afin de capitaliser sur notre savoir-faire et réutiliser rapidement les modèles. Sachant qu'une bonne réutilisation implique de bien documenter et de coder proprement. Nous n'excluons pas

de solliciter des start-up ou des PME qui ont de l'avance sur tel ou tel sujet afin qu'elles nous fournissent un réservoir algorithmique complémentaire.

Criteo

La forte contrainte du temps de réponse a-t-elle un impact sur la façon dont vous concevez les modèles prédictifs ?

Dans ce contexte de temps réel, nous sommes évidemment limités par la barrière physique. Résultat : si quelqu'un réalise un modèle plus sophistiqué mais qui s'avère plus lent, nous ne le mettrons pas en production. De même, nous évitons d'échafauder des édifices trop complexes incluant le stacking de modèles ou des produits de matrices. En fait, le gros de notre travail quotidien porte sur l'optimisation. A la base, tout modèle est optimisé pour générer des clics ou des ventes mais beaucoup d'autres paramètres importants peuvent être constamment améliorés. D'une manière générale, le choix des variables que nous utilisons a bien plus d'effets que la modification des modèles sous-jacents. La plupart des data scientists que je rencontre cherchent à élaborer des modèles très compliqués à performance élevée dans un cadre d'hypothèses donné alors que remettre en cause les hypothèses apporte souvent davantage de bénéfices.

Effectuez-vous cette recherche de nouveaux modèles en environnement de développement séparé ?

Historiquement, nous ne disposions quasiment pas de chercheurs. Tout était réalisé en environnement de production : les ingénieurs développaient directement leurs POCs (proof of concept ou preuve de faisabilité) en C#, lesquels partaient très vite en A/B test. La situation évolue. Nos chercheurs ne sont pas pour autant dispensés de respecter certaines exigences : produire un modèle très efficace mais qui ne serait pas calqué sur le volume d'événements brassé par Criteo n'a aucun intérêt. Nos modèles ont par ailleurs une durée d'efficacité très courte, tout simplement parce que l'historique tout comme les intérêts et les comportements d'achat des internautes changent. La création des modèles est automatisée et industrialisée. Nous les re-générons toutes les 4 ou 6 heures.

Safran

Comment choisissez-vous les métriques à utiliser ?

Nous commençons avant tout par identifier ce qui est interprétable, identifier les corrélations, puis nous apportons des outils de visualisation. Notre objectif n'est pas de délivrer à tout prix un modèle prédictif mais d'abord de comprendre le phénomène pour ensuite agir sur les causes.

Trouver la bonne métrique ne se fait qu'avec le retour utilisateur qui détermine si l'outil apporte de l'information supplémentaire. La phase de prototypage sur données chaudes permet de corriger une métrique inadaptée retenue en POC. Nous ne sommes pas dans une approche d'optimisation, nous apportons plutôt notre capacité à explorer les données pour aider l'expert à mieux naviguer et visualiser ses données.

En fine, plus que l'AUC (Area Under Curve), c'est l'impact des résultats obtenus en pratique grâce à l'analytics (en termes de ROI par exemple) qui est la métrique que nous suivons.

IV.1.4 – APPRENTISSAGE ACTIF

Les étapes d'évolution d'un modèle présentées précédemment supposent que le jeu de données utilisé pour l'apprentissage est statique, c'est-à-dire qu'il ne subit aucune modification dans le temps. En réalité, ce dernier est souvent amené à changer, que ce soit de manière discontinue ou de manière continue. Intégrer de nouvelles données permet souvent d'améliorer la précision et/ou capturer des comportements nouveaux (ex : les fraudeurs). De fait, il devient nécessaire de ré-entraîner le modèle (plus ou moins régulièrement) avec ces nouvelles données.

Traitement manuel / hors-ligne

Nous nous plaçons dans la situation où le réentraînement se fait manuellement et où la brique évaluation, le code externe et interne sont fixes. Autrement dit, nous exposons l'effet isolé de l'intégration de nouvelles données. On distingue alors deux cas : celui où les hyper-paramètres restent fixes pour lequel l'intégration des données a pour seul effet d'ajuster les paramètres du modèle et celui où les hyper-paramètres changent (manuellement ou après utilisation de la brique évaluation comme suite à un grid-search).

D'un point de vue algorithmique, on parle de « offline learning ». La question de la fréquence de réentraînement est primordiale. Elle dépend de la périodicité moyenne à laquelle les données à disposition changent de manière significative et de l'horizon décisionnel voulu par les décideurs, jusqu'à plusieurs fois par jour pour la publicité en ligne.

Traitement automatisé / en ligne

Quand les données sont trop volumineuses pour être traitées d'un bloc, qu'elles sont générées intrinsèquement en fonction du temps comme pour un cours boursier ou encore que le temps de traitement est supérieur au temps d'acquisition des données, il faut apprendre de manière séquentielle. Le référencement des pages Google est directement dépendant de la navigation des utilisateurs, sans intervention humaine. Désormais **les étapes suivantes sont automatisées : réentraînement du modèle, ajustement des hyper-paramètres, ajustement des paramètres et évaluation du modèle.**

En d'autres termes, l'état interne du modèle est mis à jour automatiquement. D'un point de vue technique, l'automatisation de ces tâches ne pose pas de problème particulier. **La vigilance doit être portée sur le rythme de mise à jour.** En effet, faut-il réentraîner un modèle dès qu'une quantité infime de nouvelles données est disponible ? Ou vaut-il mieux les accumuler et les traiter par batch successifs et ainsi assurer les bonnes conditions de mises à jour évoquées dans la section précédente ? Comment s'assurer que la nouvelle version n'est pas biaisée ?

.....*Abdellah Kaid-Gherbi / Long Do Cao*

IV.2 | La maîtrise de la performance du modèle de bout en bout

Mesurer l'évolution de la performance d'un modèle est crucial pour contrôler les résultats et s'assurer que la promesse émise lors des prototypes se vérifie lors du passage en pilote puis en industrialisation. Choisir une bonne métrique de performance est donc primordial pour piloter un projet de data science. Cependant, les utilisateurs d'un POC et d'un pilote étant distincts, la nature des données collectées changeante, les métriques utilisées sont souvent appelées à être adaptées.

IV.2.1 – ATTENTION AUX MÉTRIQUES DES POCS

En phase POC, on assimile souvent la performance du projet data science à la performance du modèle de Machine Learning. Ainsi, des métriques plutôt techniques sont typiquement utilisées. **L'aire sous la courbe ROC, ou AUC et le lift figurent parmi les métriques les plus fréquemment utilisées. Ces métriques sont cependant problématiques car elles peuvent induire les opérationnels en erreur** : le lift, qui compare les performances du modèle avec un choix aléatoire, risque d'être trompeur si le taux de cible est très faible. Par exemple, une valeur de lift en apparence très satisfaisante, peut impliquer la nécessité d'appeler des milliers de personnes pour trouver quelques dizaines d'acheteurs.

IV.2.2 – TRANSFORMER LA MÉTRIQUE SCIENTIFIQUE EN MÉTRIQUE MÉTIER

Le passage d'une métrique scientifique (AUC, Precision Recall) à une métrique métier intelligible pour tous et dont les évolutions ont des conséquences mesurables par tous est donc essentiel pour réussir en pilote. Une dégradation de 10 % d'une AUC n'est pas parlant pour un opérationnel, savoir qu'un modèle risque de catégoriser des clients valides en fraudeurs l'est nettement plus.

La transformation de la métrique doit donc s'appuyer sur une compréhension fine du contexte dans lequel le modèle est utilisé. Cette compréhension permet d'identifier si le contexte peut tolérer un fort niveau de faux positifs (une erreur commise par le modèle ; valable si le coût du faux positif est faible comme envoyer un email, invalidé si le faux positif se traduit par un arrêt de production d'une machine industriel). Dans le premier cas de figure, nous avons des modèles qui captent beaucoup d'événements au risque d'avoir un taux de faux positifs plus haut.

Dans le second cas de figure, le contexte exige du modèle un très fort niveau de précision : il est acceptable que le modèle rate des événements mais il ne faut pas qu'il se trompe car les impacts financiers sont conséquents.

L'équipe doit prendre en compte l'impact des décisions du modèle pour trouver un paramétrage acceptable des valeurs de seuil, puis trouver les bons indicateurs pour que les opérationnels soient en mesure de comprendre les décisions du modèle et identifier les éventuelles dérives lorsque le modèle est en production.

IV.2.3 – APPRENTISSAGE PAR RENFORCEMENT

En data science, les deux cadres d'apprentissage les plus utilisés sont le supervisé (on connaît la variable à expliquer) et le non-supervisé (on cherche à regrouper les observations ayant les mêmes caractéristiques). Un autre type d'apprentissage, dit par renforcement, désigne un ensemble de méthodes dont les règles de décision évoluent grâce à un système de récompense (positive ou négative). L'algorithme, exposé à des nouvelles données, reçoit une récompense en fonction de la qualité de sa prédiction. Dans ce contexte, inspiré à l'origine par des études de psychologie comportementale, un agent doit prendre des décisions dans un environnement inconnu, et reçoit des récompenses plus ou moins fortes sur la base de la décision prise. Petit à petit, l'agent commencera à connaître de mieux en mieux son environnement, à comprendre quelles sont les décisions les plus productives, et à les répliquer de plus en plus fréquemment. Son comportement changera donc, non pas en raison d'une règle explicite, mais plutôt sous la pression environnementale. Ce genre d'algorithme est très naturellement lié avec **le compromis entre exploration et exploitation**, car à chaque action l'agent doit choisir entre agir de manière à maximiser la récompense sur la base de sa connaissance actuelle du système ou chercher d'autres stratégies encore plus rentables (au risque de tomber sur une action mal récompensée). **Un algorithme d'apprentissage par renforcement permet donc de ne pas avoir à réentraîner un modèle, ce qui en fait une solution de choix lorsque l'on doit prendre des décisions où les stratégies évoluent rapidement au cours du temps.** Cette technique est de la même manière très utile dans les cas où le modèle doit fonctionner en temps réel, avec une mise à jour incrémentale à chaque nouvelle donnée. Un cas d'application courant est le ciblage publicitaire sur le Web, qui permet de pousser la bonne publicité au bon client.

Un exemple est l'algorithme dit « **bandit manchot** », dont le principe est équivalent à un joueur devant plusieurs machines à sous. Nous sommes face à différentes stratégies (les bandits) et nous cherchons à trouver le bras du bandit qui génère le plus de gains, et ce, le plus rapidement possible. Après avoir joué tous les bandits une première fois, l'idée est de rejouer le meilleur bras avec la probabilité p (exploitation) et un bras au hasard avec une probabilité $1-p$ (exploration). Au fur et à mesure que notre connaissance du système augmente, nous pourrons de plus en plus exploiter le bras le plus rentable.

Même si de tels modèles apprennent au fil de l'eau, leurs paramètres et les variables explicatives utilisées pour les construire doivent être modifiés au bout d'un certain temps. On retrouve alors le même dilemme qu'auparavant, à savoir un arbitrage entre exploration, exploitation et maintenance.

IV.2.4 – A/B-TESTING

Le modèle a désormais été entraîné puis testé sur un échantillon de test et nous sommes donc prêts à l'appliquer sur une population cible ; mais comment évaluer nos résultats *a posteriori* ? L'A/B-Testing est un outil très fréquemment utilisé en marketing pour évaluer les performances d'un modèle.

Le principe est de diviser aléatoirement la population visée en deux groupes, à savoir le groupe de test et le groupe de contrôle (également appelé « placebo »). Le modèle est appliqué au groupe de test, tandis que le groupe de contrôle ne subit aucune modification. Si la randomisation des deux groupes a bien été faite sans introduire de biais (on effectue en général des tests statistiques comme le test de Student pour s'assurer que les deux groupes sont bien identiques), on peut alors attribuer la différence de performance entre les deux groupes (par exemple un taux de conversion plus élevé) à l'effet du modèle sur le groupe de test. De façon très simplifiée, **on peut résumer l'A/B test à "est-il mieux de faire que de ne rien faire ?"**.

IV.2.5 – VERS DATA OPS

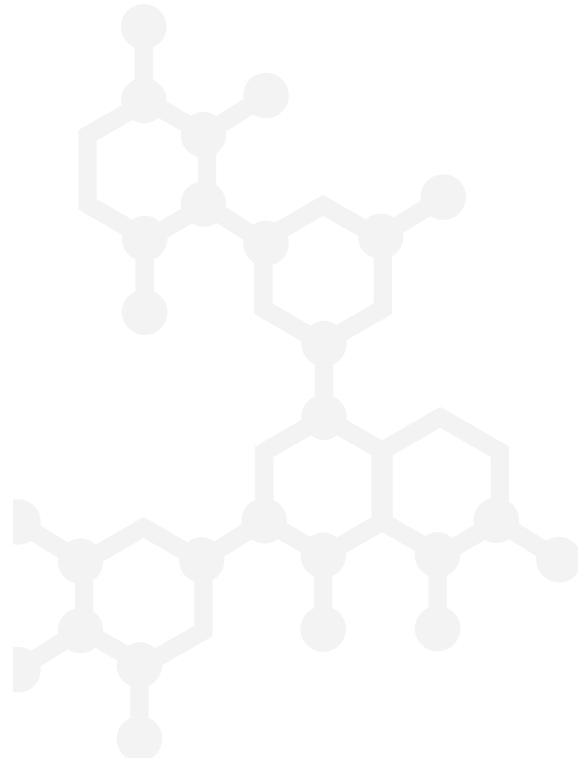
Une étape supplémentaire dans l'industrialisation des développements passe par l'application des méthodes et des outils du DevOps. Chaîne de fabrication logicielle, gestionnaire de configuration, automatisation des déploiements et des tests, utilisation de la virtualisation et des containers, tous ces composants facilitent le passage à l'industrialisation. Les projets Big Data peuvent donc reprendre tels quels ces principes et méthodes et s'appuyer sur les retours d'expérience des projets Big Data. L'objet de ce livre blanc n'étant pas de détailler le DevOps, voici néanmoins quelques points d'attention spécifiques aux projets Big Data.

Les données d'apprentissage des modèles font partie de la configuration logicielle. Avec des données différentes, un modèle aura un comportement différent. Il est donc nécessaire d'inclure dans la gestion de configuration les données d'apprentissage, les données de test et les mesures de performance associées. Ainsi, il est possible de reproduire le comportement d'une version de modèle pour un résultat donné.

Cette possibilité de versionner les données d'apprentissage n'est toutefois pas applicable lorsque les modèles sont en apprentissage automatique (traitement en ligne) ou dans un cadre d'apprentissage par renforcement. Dans ce cas, il faut positionner sur la chaîne de déploiement logiciel des tests de mesure de la qualité du modèle permettant de s'assurer que les nouvelles versions d'un même modèle ne seront pas susceptibles de dégrader la performance ou les résultats de l'application. Cette chaîne de contrôle n'existe pas telle quelle dans les outils de DevOps et reste à bâtir dans chacun des cas.

Enfin, les tests de mesure (effectués par un robot de déploiement de type Jenkins) devront intégrer des cas de tests plus subtils que les cas binaires de type JUNIT test passant / non passant, mais plutôt intégrer des valeurs de seuil paramétrables en fonction d'un

niveau de performance attendu. Dans cas aussi, les frameworks de tests connus des développeurs ne remplissent pas encore la fonction et les data scientists doivent le plus souvent construire une surcouche dédiée.



..... Alberto Guggiola / Ysé Wanono

IV.3 | Dérive des données

Une fois industrialisé, un modèle renvoie une sortie (des prévisions par exemple) en se basant sur des données d'entrée. **Ces données d'entrée peuvent dériver dans le temps pour plusieurs raisons.** Premièrement, le processus de récolte de données a imperceptiblement varié : une caméra prend un petit angle, on retire une catégorie dans un formulaire. Deuxièmement, la qualité des données diminue pour cause de mauvais fonctionnement : un capteur se met à renvoyer des valeurs absurdes. Bien sûr, un apprentissage actif utilisera les nouvelles données pour réentraîner le modèle, qui tiendra alors compte de ces variations d'autant plus que la dérive des données est lente. Cependant, la dérive des données pose la question suivante : quelles garanties possède-t-on sur la qualité des prévisions et quelles stratégies mettre en œuvre pour garder de la robustesse ?

En Machine Learning, la théorie mathématique peine parfois à expliquer les performances des modèles. Seule une **approche empirique de mesure** avec des échantillons de test est à même de donner des indications sur la robustesse d'un modèle.

IV.3.1 QUELLE VALIDATION MATHÉMATIQUE ?

Une des approches possibles du problème, exposée par **Stéphane Mallat** lors de son séminaire au **Collège de France*** est de considérer que **les algorithmes de Machine Learning sont des fonctions mathématiques fournissant des prédictions et créant des « invariants » par rapport aux données**, c'est-à-dire que pour certaines variations autour des données d'entraînement, la prédiction reste la même. Dans son exposé, Stéphane Mallat démontre en faisant une analogie avec la théorie des ondelettes pourquoi en traitement d'image, l'enchaînement d'opérations linéaires et non linéaires dans les réseaux de neurones convolutionnels fait sens, et pourquoi cela crée des « invariants » robustes à bon nombre de transformations. Il s'agit ici d'un champ d'exploration scientifique en devenir dont les résultats conditionnent le développement du Machine Learning à des usages sensibles.

IV.3.2 SIMULATION DES DÉFORMATIONS DES DONNÉES

En attendant une fondation théorique plus solide, une façon de réduire l'incertitude est de **simuler les déformations possibles des données afin d'étudier les réactions du modèle**. En traitement d'image cela est particulièrement aisé car un certain nombre de déformations sont facilement simulables (variations de luminosité, couleurs, rotations, translations). Par contre dans des espaces de dimension plus grands, les axes n'ont pas d'interprétations simples. Il faudra alors appliquer des transformations en adéquation avec le métier en faisant des hypothèses plausibles de déformation des features.

Cette méthode n'est pas entièrement satisfaisante, car il y aura toujours des déformations non testées. Elle s'applique principalement sur les variables explicatives et non les cibles. Néanmoins, elle donne une confiance supplémentaire quant à la capacité du modèle à résister à certaines déformations des données.

IV.3.3 DÉTECTION DE VALEURS ABERRANTES EN ENTRÉE

Une autre méthode est **d'utiliser un algorithme de Machine Learning de type non supervisé avant de soumettre la donnée à prédire au modèle final** afin de détecter les potentielles valeurs aberrantes et les écarter de la prédiction. Cette méthode est particulièrement adaptée pour les variables continues et donne de bons résultats lorsque les distributions des variables sont plutôt régulières. Dans ce cas il faut alors entraîner deux modèles, le modèle de prédiction lui-même et le modèle de détection des valeurs aberrantes.

IV.3.4 INTERVALLES DE CONFIANCE

Enfin, une autre approche consiste à déterminer des intervalles de prédiction, un intervalle de prédiction étant l'estimation d'un intervalle dans lequel les futures observations tomberont avec une certaine probabilité. Ainsi, si une prédiction ne tombe pas dans cet intervalle, elle sera rejetée. À noter **qu'il ne faut pas confondre cette notion avec la probabilité retournée par un classifieur**, qui n'est pas un intervalle de confiance, erreur souvent commise. En effet une probabilité de 0,5 peut aussi bien traduire le fait que l'on a appris très peu à cause d'un manque initial d'exemples d'entraînement, tout comme le fait que la décision est incertaine malgré un nombre d'exemples d'entraînement élevé. Des solutions existent pour déterminer ces intervalles, comme les Quantile Regression Forests ou la Prédition conforme, qui mériteraient d'être généralisées et surtout beaucoup plus implémentées dans les librairies standards de Machine Learning.

IV.3.5 LOGICIELS POUR TRAITEMENT AUTOMATIQUE DU MACHINE LEARNING

Afin de pallier les problèmes présentés ici, les logiciels de Machine Learning proposent maintenant une brique d'automatisation du choix d'algorithme ou de l'enchaînement hiérarchique des modèles de données. Relativement récents, ces composants seront amenés à jouer un rôle de plus en plus important dans le contrôle des résultats de Machine Learning.

..... *Vincent Dejouy*

IV.4 | Effets de rétroaction

Une des limites de la phase de prototypage est que les résultats obtenus n'ont pas eu d'effet réel sur la cible. Or, lorsqu'un modèle est mis en production, ces effets de rétroaction apparaissent. Trois effets sont à prendre en compte : la pression de sélection, l'effet systémique et la dégradation de la base d'apprentissage.

IV.4.1 – PRESSION DE SÉLECTION

Un modèle de détection de fraude apprend à détecter des potentiels fraudeurs. Il est donc facile et tentant d'extrapoler les résultats de ce modèle sur la vie réelle (si nous avions mis ce modèle en production, nous aurions économisé tant, donc le ROI du projet est X %). **Or, le modèle va avoir un effet sur les comportements de fraude.** Les fraudeurs les plus rustiques vont être détectés très vite et ne recommenceront plus, les fraudeurs les plus malins vont adapter leur comportement de manière à rendre leurs procédés moins facilement détectables et les personnes tentées par la fraude vont s'abstenir de copier les fraudeurs. Le modèle exerce donc une pression de sélection (au sens darwinien) sur la cible qui a un effet assez fort sur la performance. Dans le cas de la fraude, les performances du modèle sont très bonnes au départ puis se dégradent rapidement (sans apprentissage continu) et au bout d'un certain temps, plus aucun fraudeur n'est détecté. Cela veut-il dire pour autant que le modèle n'est plus efficace et qu'il faut stopper les investissements ? Probablement pas. D'une part, en rendant la fraude plus complexe, le modèle diminue le nombre de cas simples (ce qui correspond à la dégradation de la performance) et a un effet dissuasif (ce qui correspond à des fraudeurs potentiels dissuadés). La quantification de ces deux effets : diminution des cas détectés et fraudes évitées n'est pas évidente et a pourtant un effet significatif sur les retours sur investissement.

Il ne faut donc pas avoir une vision trop simplificatrice du problème (extrapoler directement les résultats du POC) ni enfermer le projet dans une vision de performance trop étroite.



IV.4.2 – EFFET SYSTÉMIQUE

Un autre effet des modèles sur la réalité doit être pris en compte, celui d'**effet systémique**. Dans une situation de POC, le modèle est extérieur au système qu'il étudie et n'a donc pas d'effet sur son comportement. **En production, les décisions prises sur la base des résultats du modèle modifient le comportement du système et la nature des données émises.** Notre modèle est donc devenu partie intégrante du système qu'il supervise (alors que précédemment il était extérieur) : il s'agit d'un effet systémique.

Un modèle de prévention de panne dans les transports s'appuie sur les données émises par des objets communicants. Ayant une base d'apprentissage consolidant les signaux émis par l'électronique embarquée, le modèle apprend à reconnaître des séquences de messages annonciatrices de pannes futures. Une fois mis en production, le modèle évite donc l'occurrence des pannes qu'il a appris à détecter : les séquences d'apprentissage n'apparaissent donc plus puisque l'événement est évité ! Or, il est essentiel de continuer à faire apprendre le modèle sur les données réelles (notamment dans le cas de l'électronique embarquée, la forme et la fréquence des messages changent au rythme des mises

à jour logicielles et des paramétrages qui ne sont pas maîtrisés par le modèle). La performance du modèle va donc inexorablement chuter. Il existe heureusement une manière d'éviter cet effet : il faut labelliser la panne évitée a posteriori, valider la qualité de la prédiction et renseigner la base d'apprentissage avec la séquence d'événements associés (avec un label de type panne évitée confirmée). De la sorte, la base d'apprentissage reste suffisamment riche pour que l'effet systémique soit limité. En revanche, cela signifie qu'il faut recourir à un expert humain qui, en aval du modèle, assurera cette labellisation et l'alimentation de la base d'apprentissage.

Le modèle en production ne peut pas vivre tout seul et modifie la réalité qu'il observe.

IV.4.3 – DÉGRADATION DE LA BASE D'APPRENTISSAGE

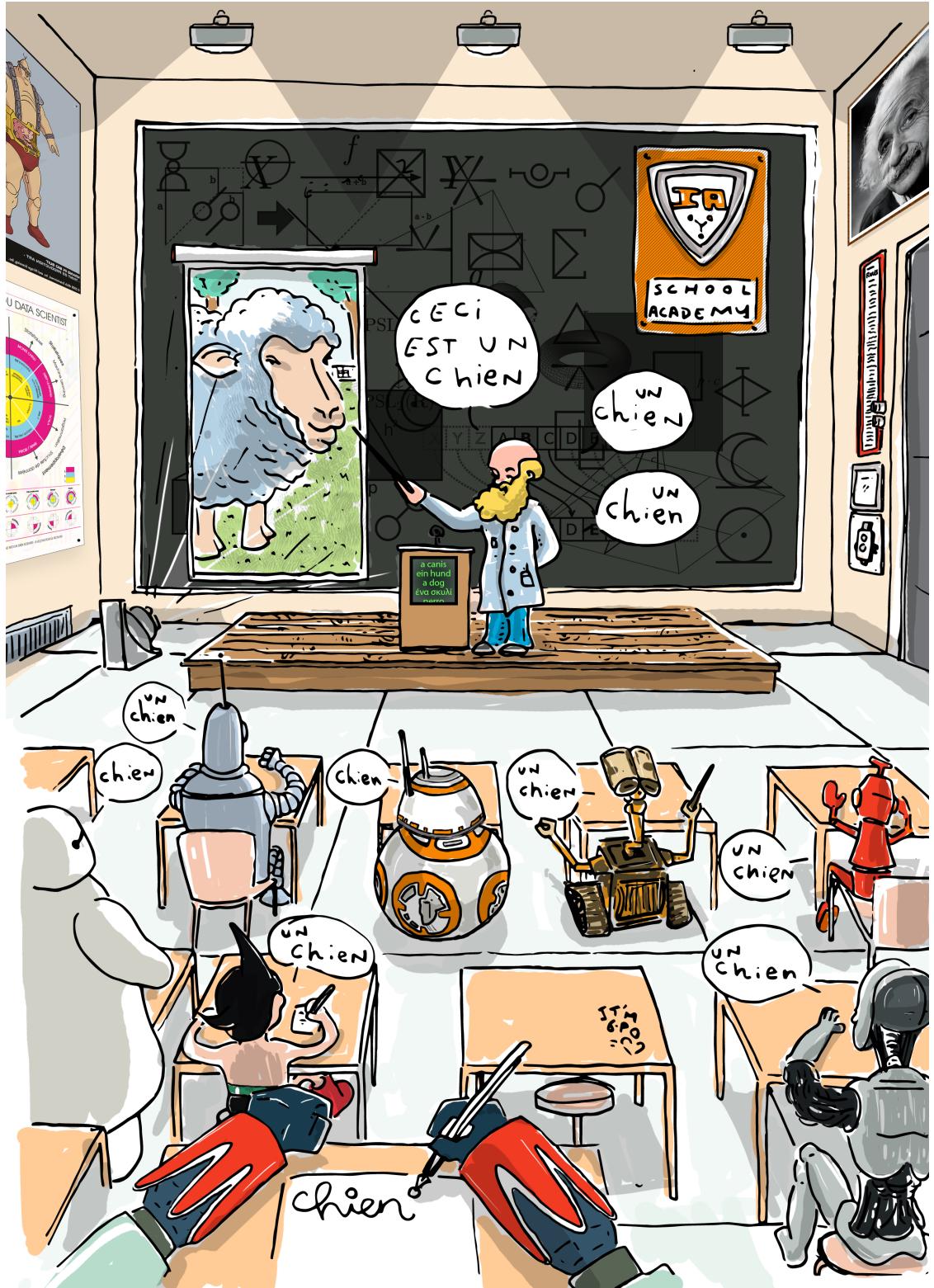
Un dernier effet à prendre en compte en production est **la disparition de certains événements de la base d'apprentissage au fur et à mesure de l'exploitation du modèle**. Les faux positifs dans un cas de fraude (les personnes détectées à tort comme fraudeurs) peuvent constituer une masse significative de manque à gagner (des contrats refusés) qui détruit la rentabilité de la démarche. Systématiquement rejetés à tort, le système ne sait plus les détecter comme des non-fraudeurs et perd la base d'apprentissage qui y est associée.

La meilleure solution consiste à maintenir une partie du flux entrant hors du modèle de prédiction afin de pouvoir le réintroduire dans la base d'apprentissage sans détérioration due au modèle.

Le modèle peut se comporter comme un exagérateur d'un phénomène observé, ce qui pose des problèmes de performance.

Au-delà de la dégradation de cette base, la dépendance d'un modèle à sa base d'apprentissage rend le Machine Learning propice au hacking et au détournement : **il suffit de fausser l'apprentissage pour rendre une intelligence artificielle parfaitement idiote.**

..... Emmanuel Manceau



CONCLUSION :

Vers la transformation par les algorithmes.

Nous sommes convaincus que le cheminement vers l'industrialisation du Big Data, organisé avec méthode et rigueur, sera suivi avec succès par la grande majorité des entreprises. Pour autant, cette transformation aura des effets profonds sur l'organisation des entreprises et leurs écosystèmes, sous l'œil avisé du législateur.

Les relations entre directions informatiques, directions métiers, éditeurs de logiciels, opérateurs cloud et sociétés de services évoluent avec un mouvement de balancier très fort en faveur d'entités capables de porter les projets sur la totalité de leur cycle de vie.

L'internalisation de compétences techniques, les partenariats d'*open innovation* et l'acceptation des cycles d'*essai-erreur* propres à l'innovation vont à rebours des modes d'organisation que nous avions observés jusqu'ici. Cette transformation des cultures des grands groupes est d'ores et déjà lancée et il ne se passe plus une semaine sans qu'un grand cabinet de conseil y fasse allusion le plus souvent sous la forme d'*agilité*.

Des structures hybrides associant des start-up et des PME innovantes à la valorisation des données des grands groupes vont continuer à se développer et renforcer les partenariats d'*open innovation*. Les mieux placés auront compris que leur compétitivité viendra aussi du dynamisme de leur écosystème de partenaires. Le mot de sous-traitants aura d'ailleurs disparu.

Jusqu'ici seul au monde, le data scientist prend conscience de ses limites et les plus avisés ont déjà compris que les directeurs de projets, les architectes et les consultants en organisation auront leur carte à jouer dans la réussite des projets.

L'actuel foisonnement des technologies regroupées sous l'appellation Big Data ne doit pas masquer le fait que l'*open source* est pour la première fois un précurseur. **Linux** est né après **Unix**, **JBOSS** après les serveurs d'application **J2EE**, mais **Hadoop** ou **scikit-learn** n'ont pas d'équivalents dans le monde commercial. De leur côté, les opérateurs de plate-forme (IAAS, PAAS) misent sur l'Internet des objets pour proposer des solutions intégrées de bout en bout. Des alliances jusqu'ici inattendues entre grands groupes dits traditionnels (PSA, Renault ou encore SANOFI) et les opérateurs de plateforme (Microsoft, IBM, Google) vont remettre en cause les fournisseurs traditionnels de la direction informatique.

La nature particulière des modèles de Machine Learning rend parfois délicate sa bonne compréhension et peut mener à des actions inappropriées. Le lien fondamental entre le code, les données d'apprentissage et les mesures de performance appelle à de nouvelles formes de forges logicielles capables d'évaluer en temps réel la performance des modèles et les déployer.

Toutefois, rien ne peut se faire sans la capacité d'une entreprise à procéder à une ré-allocation massive de ses ressources vers l'industrialisation du Big Data. Sans cela, il est probable que les organisations soient condamnées au POC permanent et se fassent petit à petit grignoter leur avantage compétitif ou leur clientèle. Ces décisions lourdes d'investissement prennent encore du temps et constituent aujourd'hui le principal frein à l'industrialisation.

Une fois toutes ces étapes franchies, nous aboutirons alors à une nouvelle étape dans la transformation digitale. Lorsque toutes les organisations auront acquis la maîtrise du processus d'industrialisation, l'enjeu va se déplacer vers une meilleure valorisation des données.

Cette étape supposera l'emploi d'algorithmes de plus en plus performants et complexes. Ce processus de transformation algorithmique commence à se dessiner pour les acteurs les plus avancés et viendra petit à petit se diffuser à l'ensemble des entreprises. Choisir les bons modèles, étudier comment les combiner, mieux interpréter les résultats ou encore automatiser plus fortement certains processus avec des composants adéquats ; ces enjeux décrivent les futurs challenges de transformation algorithmique.

Certains en ont déjà essuyé les plâtres comme Facebook avec les polémiques nées de son système de recommandation d'articles ou encore le monde de la finance algorithmique jugé responsable d'une augmentation de la volatilité et des « flash crash ».

Mais on peut dès maintenant percevoir la plus-value du bon usage d'un algorithme. Criteo tire son avantage compétitif d'un meilleur taux de transformation des publicités ciblées ; Amazon a fondé son succès sur un des meilleurs taux de conversion du Web, un opérateur téléphonique économise des frais en détectant au plus tôt des fraudeurs potentiels.

Demain, un industriel tirera un avantage à produire un équipement qui tombera moins en panne, un distributeur d'une meilleure rotation des stocks de sa supply chain ou encore une administration d'une meilleure gestion de ses ressources humaines. Toutes ces améliorations exploiteront des algorithmes et des données dans un environnement industrialisé.

La récente annonce faite par Barack Obama, prévoyant un investissement de quelque 80 milliards de dollars dans l'intelligence artificielle, devrait finir de convaincre de l'importance de cette transformation. Pour reprendre le titre d'un livre de l'essayiste Nicolas Bouzou : « On entend l'arbre tomber mais pas la forêt pousser », les leaders de demain innovent dès aujourd'hui sur leurs algorithmes dans le secret des laboratoires et construisent patiemment leur futur avantage concurrentiel.

..... *Emmanuel Manceau*

Notes

AVEC LA PARTICIPATION DE :

Jeremy Harroch, Guillaume Perrin-Houdon, Emmanuel Manceau, Long Do Cao,
Gill Morisse, Matthieu Vautrot, Ysé Wanono, Stéphane Jankowski, Nicolas Gibaud,
Abdellah Kaid Gherbi, Youssef Benchekroun, Alberto Guggiola, Vincent Dejouy,
Jean-Matthieu Schertzer, Robin Lespes, Michèle Marchand.

GRAPHISME ET ILLUSTRATIONS

Aurélien Gomez
Aureliengomez@gmail.com
neografix-creation.com

IMPRESSION & BROCHAGE

SCRIPT LASER
29, boulevard Malesherbes
75008 Paris

