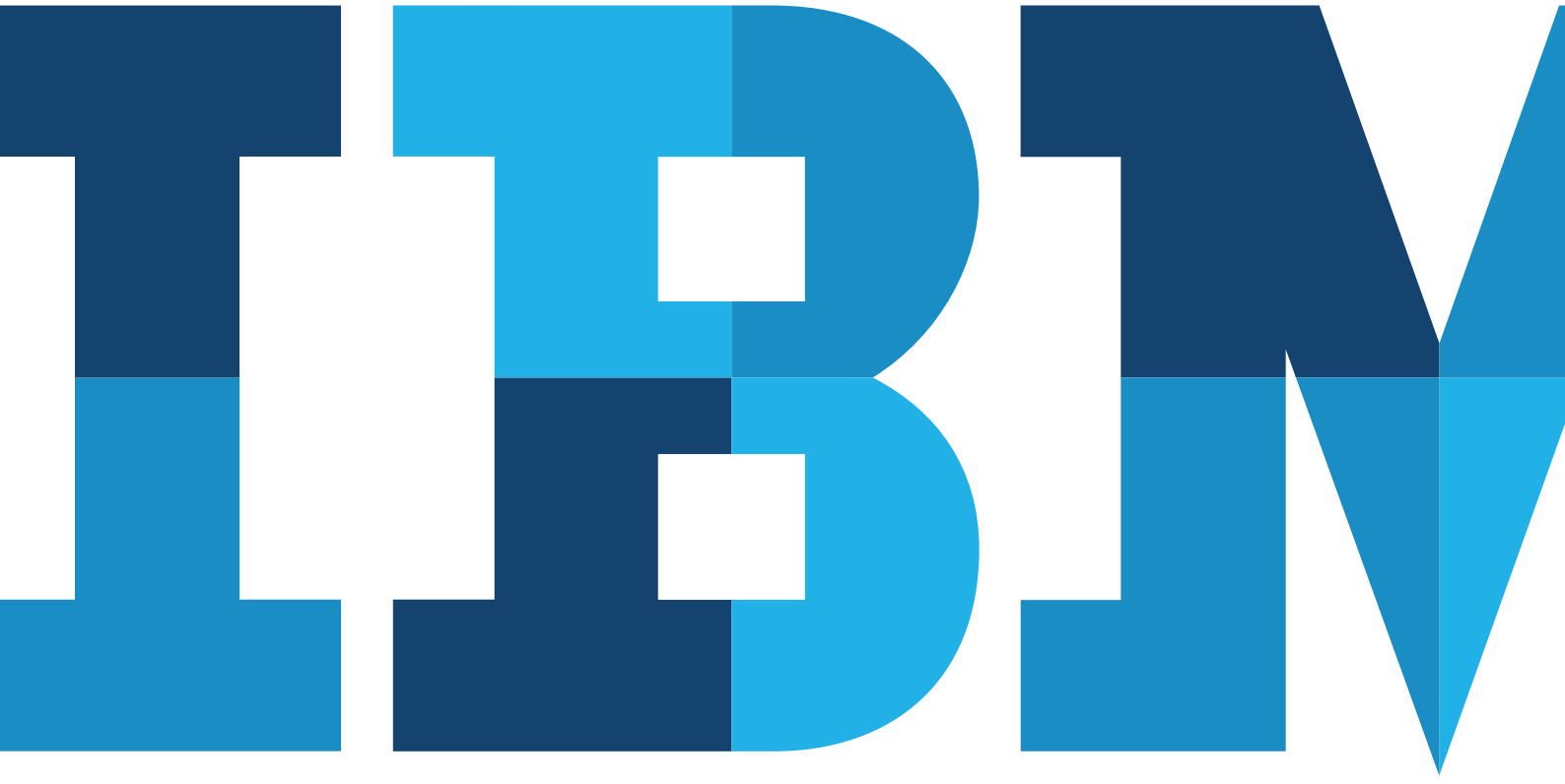
IBM Global Business Services
Livre blanc

Février 2017

Data Science for Business

Le Data Science ou comment extraire l'information clé à partir de volumes de données massifs.



IBM

Sommaire

2 Qu'est-ce qu'un Data Scientist ?

5 Quels sont les projets types de Data Science ?

7 L'industrialisation du Data Science

Synthèse

Quel que soit leur secteur d'activités, les entreprises continuent à accumuler des masses de données, mais recourir au Data Science (littéralement la « science des données ») pour extraire efficacement la substance de l'information n'est pas une mince affaire. Pour relever ce défi, il faut constituer une équipe diversifiée d'experts en données, mettre en œuvre une méthodologie de projets éprouvée et bien entendu choisir la méthode de déploiement adéquate. Ce livre blanc montre précisément aux décideurs comment ils peuvent augmenter leur performance opérationnelle grâce à une approche Data Science informée.

Qu'est-ce qu'un Data Scientist ?

Le terme « Data Scientist » (scientifique des données) n'est pas nouveau. Il a été inventé en 2008 par D.J. Patil et Jeff Hammerbacher avant d'être repris par les analystes respectifs de LinkedIn et Facebook[1]. Neuf ans plus tard, personne ne sait toujours en quoi consiste réellement le rôle du Data Scientist. Si vous les interrogez, la plupart d'entre eux vous diront que les Data Scientists ne sont pas des magiciens (ou alors ils sont extrêmement rares). Lorsqu'on considère le profil de compétences clés d'un Data Scientist (science informatique, mathématiques, statistiques, expertise métier), c'est loin d'être étonnant.

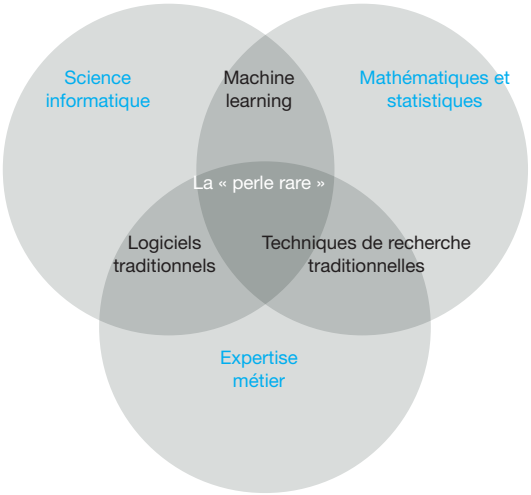


Figure 1 : Profil de compétences clés d'un Data Scientist.

Le Data Scientist qui prétend être excellent dans tous ces domaines est soit un génie, soit quelqu'un qui n'a pas compris la dimension du terme « virtuose ». Même si certains Data Scientists sont manifestement des experts en mise en œuvre de plateformes Big Data et en codage Pig, Hive ou Scala, il y a fort à parier qu'ils soient moins calés en techniques mathématiques et statistiques élaborées, ou qu'ils aient une connaissance limitée du ou des secteurs dans lesquels s'inscrit leur travail. Peut-être serait-il plus judicieux de se concentrer sur les membres d'une l'équipe de Data Science, qui s'attachera davantage à regrouper l'ensemble des compétences requises en Data Science plutôt que d'essayer vainement de dénicher la « perle rare » qui réunira les compétences tant convoitées.

On peut identifier trois rôles principaux autour du Data Science :

Le Data Scientist

Ici, le mot clé est « scientist » (scientifique). Le Data Scientist possède des connaissances avancées en mathématiques, en statistiques et en techniques de recherche. Il est capable de travailler directement avec l'équipe fonctionnelle (autrement dit la direction ou le département qui a fait appel à ses services) afin de délimiter ou préciser le périmètre de la problématique métier et d'en déduire une ou plusieurs expériences « data » (données). Il maîtrise parfaitement les différentes techniques de modélisation applicables à chaque problème donné (segmentation, classification, association, boosting, bagging) et est capable de concevoir une expérience qui combinera l'ensemble de ces techniques au sein d'une approche logique dans le but de résoudre la problématique métier.

Le Data Scientist a également une parfaite connaissance des langages de programmation ou des logiciels statistiques utilisés en Data Science (R, Python, SPSS et SAS, par exemple). Outre les compétences techniques, il doit également maîtriser les outils de visualisation et savoir faire parler les données. Le Data Scientist ne doit pas seulement être capable de démontrer la fiabilité d'une analyse, mais également convaincre les interlocuteurs « non techniciens » de la valeur d'une analyse/d'un modèle et justifier pour quelles raisons ils doivent l'intégrer à leurs systèmes ou processus de gestion.

Le Data Engineer

Le Data Engineer est davantage axé sur la configuration et intervient directement sur l'infrastructure de données (bases de données, clusters Hadoop, analyses continues). Cela ne signifie pas qu'il est incapable de développer des modèles prédictifs, mais qu'il travaille davantage à l'élaboration de la plateforme de données à partir de laquelle les expériences de Data Science pourront être déduites et réalisées. Le Data Engineer est un expert dont le travail consiste à identifier et extraire les données les plus pertinentes à partir d'une variété de sources, en les organisant au sein d'une structure cohérente et en effectuant une analyse initiale – voire en les agrégeant si nécessaire. En ce qui concerne le déploiement des modèles, il travaille en étroite relation avec le Data Scientist afin de mettre en œuvre le modèle retenu, parfois en recodant une partie du modèle en vue d'optimiser ses performances en environnement de production.

Le Data Analyst

Doté de solides bases en mathématiques, en statistiques et en science informatique, le Data Analyst a une connaissance approfondie du secteur, du domaine concerné et des données associées. Il aide le Data Scientist à comprendre les données et lui montre de nouvelles pistes d'investigation et, surtout, la manière dont l'élaboration de modèles peut être incorporée aux processus de gestion. Il aide également les Data Scientists à effectuer certaines tâches d'analyse élémentaires (nettoyage des données, création de fonctionnalités, analyse des tendances, KPI, etc.) et est en charge du modèle de données sur lequel est fondé le projet de Data Science. On les appelle parfois les « Data Scientists citoyens ».[2]

Une bonne équipe de Data Science entretient un bon équilibre entre ces trois profils de compétences. En outre, le nombre et la proportion des rôles (et individus) recherchés dépendent du projet à conduire. Par exemple, un projet de validation de principe (POC, *Proof Of Concept*) sera essentiellement effectué par des Data Scientists dont l'approche est davantage orientée sur l'expérimentation. Néanmoins, une fois que le concept aura été validé, et que le déploiement et l'intégration du modèle aux systèmes et aux processus de gestion auront été décidés, ce sont le Data Engineer et l'analyste qui prendront la main.

Le Data Scientist poursuivra ses activités de développement et continuera à perfectionner le modèle identifié lors de la phase POC jusqu'à la mise en production, alors qu'il serait invraisemblable de voir un Data Analyst ou un Data Engineer participer activement à la définition du concept vu son manque d'expérience probable en mathématiques et en techniques de recherche.

Bien entendu, une personne peut également endosser plusieurs rôles au sein de l'équipe de Data Science, mais il y a des chances que sa sphère de compétences soit limitée dans certains domaines. L'équipe de Data Science devra également interagir avec les personnes occupant des rôles de production plus classiques (tels que l'architecte des données), en particulier lors du passage de la phase d'expérimentation à la mise en œuvre et à l'intégration aux processus de gestion.

Une équipe de Data Science se distingue par son étonnante combinaison de compétences qui collaborent dans un but commun : comprendre et extraire des informations exploitables, à partir de vastes et complexes sources de données structurées et non structurées.

Quels sont les projets types de Data Science ?

Alors que nous cherchons à définir le profil de compétences clés du Data Scientist, il est important de comprendre également ce qui distingue un projet de Data Science d'un projet opérationnel ou IT classique.

Tout d'abord, il faut noter qu'il est (ou du moins qu'il devrait être) rare d'entreprendre un projet de Data Science seul. Les capacités fonctionnelles que peuvent apporter les Data Scientists (classification, segmentation, prédiction, prévision, etc.) doivent s'inscrire dans un contexte commercial et technologique plus vaste, et doivent résoudre une problématique ou un défi métier identifiable. Souvent, les modèles statistiques développés au cours d'un projet de Data Science sont intégrés à une solution plus vaste qui utilise les résultats de la modélisation pour aider les entreprises à acquérir une meilleure connaissance de leurs utilisateurs ou clients. Cela dit, il convient également de souligner que les projets de validation de principe (POC) et de justification de valeur (POV, *Proof Of Value*) constituent un point de départ commun pour déterminer l'efficacité d'un projet de Data Science. Généralement, ces projets constituent un coût d'investissement initial visant à examiner une problématique ou un aspect métier particulier, un domaine où les méthodes de Data Science sont censées apporter de la valeur ajoutée. Dans ces types de projets, on mettra plutôt l'accent sur les tâches d'analyse et de modélisation, même s'il faudra néanmoins y ajouter un élément métier qui définira le cas d'utilisation et le dossier commercial pour s'assurer que les modèles développés pourront avoir une incidence réelle, une fois qu'ils seront déployés au sein de l'entreprise. Un objectif et un dossier commercial clairement définis sont essentiels pour améliorer les chances de réussite d'un projet de Data Science. À défaut, l'analyse ou le modèle risque de ne pas cibler directement la problématique métier ou de ne pas apporter suffisamment de bénéfices à l'entreprise.

Lorsqu'une entreprise envisage un projet de Data Science, elle ne doit pas perdre de vue deux des qualités qui caractérisent le Data Scientist : sa curiosité et sa capacité à enquêter sur des problèmes déstructurés. Ces qualités sont la clé du succès des types de projets auxquels participe régulièrement le Data Scientist qui, en général :

- comportent un certain nombre de problèmes inter-reliés et complexes à examiner,
- incorporent d'importants volumes de données structurées ou non structurées qui sont souvent de qualité inégale,
- ne possèdent pas nécessairement un cadre ou un objectif clairement définis.

Au vu de ces facteurs, il n'est pas surprenant que les projets de Data Science ne soient pas linéaires. Il est pratiquement impossible de déterminer au début d'un projet de Data Science quels algorithmes (et paramètres associés) produiront le meilleur résultat pour un problème donné, car les résultats dépendent avant tout des données (et en particulier du volume disponible, de la diversité des variables et fonctions, ainsi que de leur qualité sous-jacente). Si l'expérience nous permet de définir différentes approches à tester, nous sommes néanmoins incapables de savoir précisément quelles variables combinées à tels algorithmes permettront d'obtenir le résultat recherché. C'est pourquoi nous procédons à des expérimentations, des évaluations et des itérations jusqu'à atteindre un résultat suffisamment satisfaisant pour pouvoir le présenter ou le déployer (selon le besoin métier préalablement défini). En d'autres termes, les projets de Data Science suivent naturellement une méthode agile (agile au sens du développement logiciel[3]). Les techniques de test et d'apprentissage s'intègrent également naturellement aux projets de Data Science, tandis que les hypothèses plausibles et parfois divergentes sont évaluées en parallèle afin d'établir le modèle optimal.

D'après un sondage KD Nuggets réalisé en 2014, CRISP-DM (CRoss-Industry Standard Process for Data Mining)[4] serait la méthodologie la plus couramment utilisée dans les projets de Data Mining et de Data Science. C'est une excellente méthodologie qui est très utilisée au sein de projets de Data Science dans tous les secteurs ; sa structure est clairement définie et facile à suivre, tandis que sa souplesse permet de procéder à des itérations entre les étapes du processus (s'il y a lieu).

Ce livre blanc n'a pas pour but de décrire en détail la méthodologie CRISP-DM qui est abordée amplement dans de nombreuses publications[5], mais les points importants à retenir sont les suivants :

- Les données constituent la clef de voûte de chaque étape du processus CRISP-DM.
- Dans un projet de Data Science, la majeure partie du temps passé est généralement consacrée à la préparation des données, car la modélisation est tout simplement inutile si les données sont incompréhensibles et de médiocre qualité. En fait, les problèmes de données peuvent très bien être découverts lors de la phase de modélisation, ce qui peut nécessiter une reprise du travail de préparation des données.
- Bien entendu, certains projets de Data Science ne seront pas nécessairement déployés et intégrés au sein de systèmes IT, mais les résultats demeureront néanmoins utiles et doivent être partagés avec les décideurs concernés.

Autre alternative à la méthodologie CRISP-DM, Standard Methodology for Analytical Models[6] basée sur CRISP-DM permet de décrire le degré d'intégration entre le processus de conception de modèles analytiques et un contexte davantage orienté métier.

Il est important que les personnes travaillant sur un projet de Data Science ne perdent pas de vue le fait que l'objectif du projet est la production d'informations exploitables pour l'entreprise.

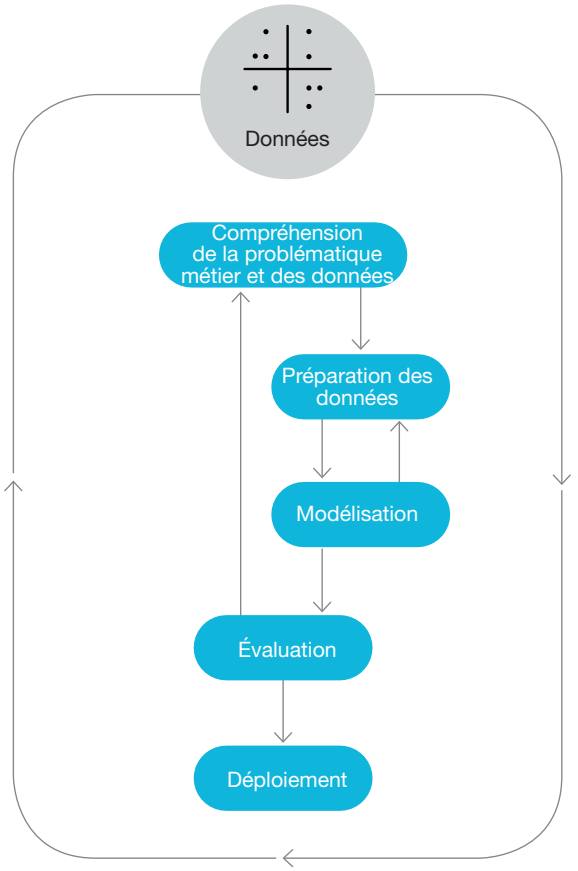


Figure 2 : La méthodologie CRISP-DM, une démarche itérative pour le Data Science.

Source : IBM GBS

L'industrialisation du Data Science

Une fois que l'analyse exploratoire initiale a été effectuée et qu'un cas d'utilisation a été retenu afin d'être examiné, on comprend aisément de quelle manière les Data Scientists peuvent capitaliser les puissantes méthodologies qui sont à leur disposition pour conduire un projet de Data Science. Et si votre projet de Data Science se composait en réalité d'une multitude de sous-projets ou cas d'utilisation à examiner ? Votre plan d'action dépendra essentiellement de la taille et de la structure de votre équipe de Data Science. En effet, plus votre équipe est importante, moins vous aurez de difficultés à la subdiviser pour effectuer plusieurs tâches en parallèle. Toutefois, en supposant que votre équipe de Data Science est de taille relativement petite (comme c'est le cas pour de nombreuses entreprises), une solution consiste à répartir le projet en une série de sprints, un sprint correspondant à une période de temps établie pendant laquelle vous devez résoudre un problème clairement défini et bien circonscrit.

L'idée est qu'à la fin du sprint vous aurez effectué plusieurs itérations des étapes Préparation des données, Modélisation et Évaluation de la méthodologie CRISP-DM, et que vous disposerez d'un modèle ou d'une analyse capable de produire des résultats suffisamment satisfaisants pour qu'ils puissent être évalués par les opérateurs (pas des Data Scientists, mais des experts métier ou des analystes de données). Peut-être que les résultats ne seront pas optimaux et que certains éléments devront être examinés plus en profondeur, mais peu importe : ces activités s'ajouteront simplement au *backlog* du prochain sprint. Le point important est que ces activités permettent de cadrer le projet de Data Science de manière à ce que l'entreprise puisse accéder à l'information réelle de manière plus rapide : en utilisant directement les analyses ou en déployant les modèles au sein de ses systèmes IT existants afin de renforcer les processus de gestion.

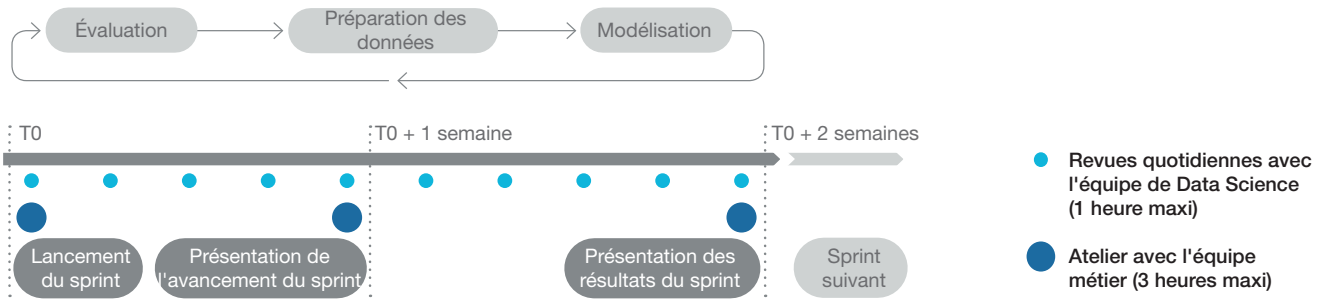


Figure 3 : l'approche Data Science pour organiser les sprints d'un projet.

Source : IBM GBS

L'objectif ultime de la méthode agile est d'accélérer la mise en œuvre des informations extraites au sein des processus de gestion, mais une simple expérimentation ne garantit pas que les informations et les modèles seront effectivement déployés ; il faut également réfléchir à la manière dont on peut passer d'un sprint Data Science au déploiement. Plusieurs méthodes existent, et la meilleure option dépendra des ressources disponibles, ainsi que de la capacité et de la volonté de changement de l'entreprise. On peut néanmoins distinguer deux options principales :

Déploiement progressif

Sur la base d'un déploiement progressif, l'idée est de définir soit une période de temps soit un nombre de sujets à étudier. Au cours de la phase de conception générale, les contraintes temps et le périmètre du projet peuvent être combinés afin de déterminer le nombre de sprints nécessaires pour mener la phase d'expérimentation du projet.

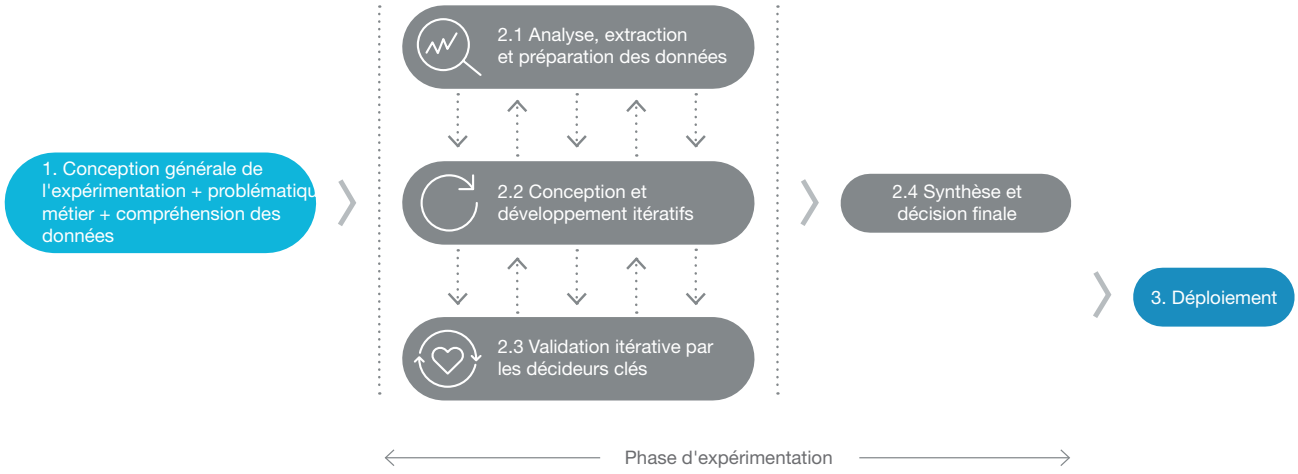


Figure 4 : Une approche progressive pour déployer les informations et les modèles issus du Data Science.

Source : IBM GBS

Il ne faut pas oublier non plus que les sprints Data Science seront probablement pris en charge par un flux de travail distinct dans le but d'extraire, de consolider et de préparer les données nécessaires au projet de Data Science (un rôle habituellement endossé par un Data Engineer). De surcroît, tout au long de l'expérimentation, l'équipe de Data Science sera assistée par les décideurs clés (experts métiers) qui aideront les Data Scientists à donner un sens aux données, mais également à évaluer les résultats de chaque sprint.

À l'issue de la phase d'expérimentation, l'équipe devra veiller à produire un dossier de déploiement convaincant dans des termes « non techniques », mais orientés métier. Souvent, l'une des principales difficultés liées au déploiement de modèles statistiques n'est pas tant leur valeur, mais leur complexité et la capacité des non-statisticiens à pouvoir les comprendre (et, inversement, la capacité des Data Scientists à pouvoir leur expliquer). C'est précisément là qu'entrent en jeu les qualités du Data Scientist à interpréter visuellement et à faire parler les données.

Une fois ce défi relevé, l'équipe de Data Science pourra commencer à établir les étapes nécessaires pour déployer le modèle ou l'analyse. L'équipe en charge du déploiement initial sera très probablement une évolution de l'équipe de projet présente lors de la phase d'expérimentation (avec une plus grande contribution des Data Engineers, ainsi que des architectes données et solutions), mais cette tâche pourrait également être confiée à un prestataire de services techniques existant si l'équipe ne dispose pas des compétences techniques suffisantes concernant les langages de programmation et les outils utilisés par la solution Data Science.

Déploiement agile

Le déploiement agile est similaire au déploiement progressif, car il débute par une phase de conception générale (parfois appelée « sprint 0 ») au cours de laquelle sont définis le nombre et le périmètre des sprints. Néanmoins, l'objectif ultime est de parvenir à des analyses ou des modèles déployés le plus rapidement possible et intégrés aux processus de gestion existants : autrement dit aboutir à un produit minimum viable.

Pour y parvenir, le projet doit être organisé d'une manière légèrement différente d'un projet de Data Science classique (qui tend à suivre un style de déploiement progressif). Soit l'équipe devra être habilitée à prendre les décisions de déploiement d'une analyse ou d'un modèle, soit l'équipe devra être élargie en y intégrant les décideurs eux-mêmes. Ce point est d'autant plus critique, qu'à la fin de chaque sprint, il faut prendre une décision afin de savoir si :

- Le sprint est achevé et ne nécessite aucun déploiement ou analyse de suivi.
- Le temps alloué au sprint doit être prolongé.
- Les extrants du sprint sont prêts à être déployés dans le cadre d'une analyse récurrente/d'un rapport. Cela s'applique aux sprints qui ne sont pas destinés à la production de résultats ayant un impact direct sur un processus de gestion ou un système IT. Le déploiement est, de ce fait, généralement assez simple.
- L'analyse ou le modèle est prêt à être incorporé(e) aux systèmes IT et aux processus de gestion existants.

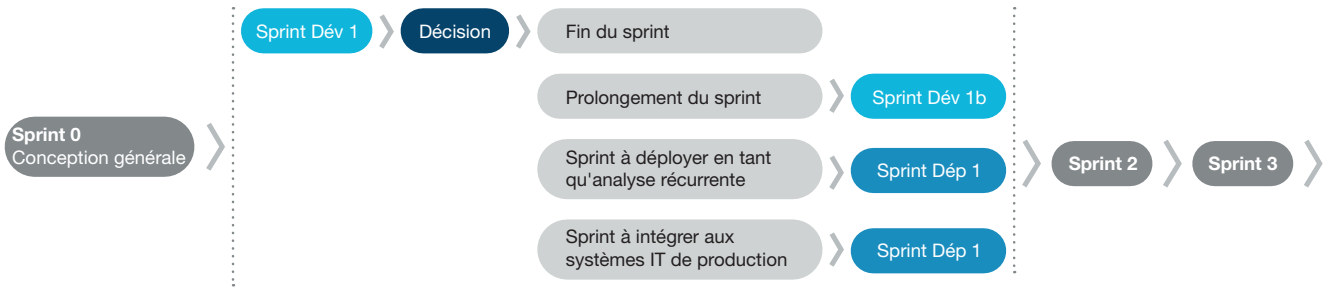


Figure 5 : Une approche agile pour déployer les informations et les modèles issus du Data Science.

Source : IBM GBS

Il est également possible d'organiser l'équipe de Data Science de manière à ce que l'expérimentation et le développement, ainsi que le déploiement soient effectués en parallèle par des équipes distinctes travaillant chacune sur l'un de ces deux aspects. Cela peut également s'avérer plus efficace dans le sens où les Data Scientists pourront se concentrer davantage sur l'expérimentation et confier les responsabilités de déploiement à une équipe de Data Engineers et de Data Analysts. Bien entendu, cela dépendra de la taille, du périmètre ainsi que l'objectif de délai de mise sur le marché du projet.

La méthode décrite dans ce livre blanc peut s'appliquer à tous les secteurs et domaines, sous réserve que le volume et la qualité des données soient suffisants pour se prêter à un projet de Data Science.

Les données jouent un rôle clé et sont l'une des raisons pour lesquelles les projets de Data Science sont agiles de nature. Néanmoins, il ne faut pas restreindre le Data Science à l'expérimentation (ou à l'étape de validation de principe). Les données sont un atout que les entreprises ne doivent pas ignorer, et elles disposent aujourd'hui de multiples méthodologies et ressources pour les exploiter pleinement.

À propos de l'auteur

Rob Worsley

Consultant
Advanced Analytics Leader, France
IBM Global Business Services (GBS)

Références

1 Thomas H. Davenport et D.J. Patel, *Data Scientist: The Sexiest Job of the 21st Century*, Harvard Business Review (octobre 2012) hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

2 Bernard Marr, *How the Citizen Data Scientist will Democratize Big Data*, Forbes (avril 2016) forbes.com/sites/bernardmarr/2016/04/01/how-the-citizen-data-scientist-will-democratize-big-data/#447c1c394557

3 The Agile Alliance, Agile 101 agilealliance.org/agile101/ what-is-agile

4 George Piatetsky, CRISP-DM, *still the top methodology for analytics, data mining, or data science projects*, article de KD Nuggets (octobre 2014) kdnuggets.com/2014/10/crisp-dm-top-methodology- analytics-data-mining-data-science-projects.html

5 Documentation IBM SPSS sur CRISP-DM ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf

6 Olav Laudy, *Standard methodology for analytical models*, Wikipédia olavlaudy.com/MediaWiki/index.php?title=Standard_methodology_for_analytical_models

© Copyright IBM Corporation 2017

IBM Corporation

Global Business Services
Route 100
Somers, NY 10589

Produit aux États-Unis
Février 2017

IBM, le logo IBM et ibm.com sont des marques d'International Business Machines Corporation, enregistrées auprès de nombreuses juridictions dans le monde. Les autres noms de produits et de services peuvent être des marques d'IBM ou d'autres sociétés. La liste actualisée des marque d'IBM est disponible sur Internet dans la section « Copyright and trademark information » à l'adresse www.ibm.com/legal/copytrade.shtml.

Le présent document contient des informations qui étaient en vigueur et valides à la date de la première publication, et peut être modifié par IBM à tout moment. Les offres ne sont pas toutes distribuées dans tous les pays dans lesquels IBM exerce son activité.

LE PRÉSENT DOCUMENT EST LIVRÉ « EN L'ÉTAT » SANS GARANTIE D'AUCUNE SORTE, EXPLICITE OU IMPLICITE, Y COMPRIS TOUTE GARANTIE D'APTITUDE À L'EXÉCUTION D'UN TRAVAIL DONNÉ ET TOUTE GARANTIE OU CONDITION DE NON-CONTREFAÇON. Les produits IBM sont garantis selon les conditions générales des contrats avec lesquels ils sont fournis.

Recyclable

GBW03367GBFR-00