Big Data : les histoires d'Hadoop finissent mal (en général)

« *Y-a-t-il une vie après les POC ?* ». Volontairement provocateur, le titre d'un petit fascicule que distribuait le cabinet de conseil Quantmetry sur le salon Big Data, qui se tenait début mars à Paris, résume assez bien les questions que se posent nombre d'entreprises face à leurs initiatives Big Data : pourquoi ces scénarios, si prometteurs dans le cadre de prototypes, ont-ils tant de mal à passer en production ? En réalité, cette question, qu'on peut décliner selon plusieurs axes (technique, organisationnel…), renvoie aux promesses irréalistes véhiculées par l'offre pour vendre le Big Data aux entreprises. « *Beaucoup d'études réalisées par des cabinets de conseil en stratégie ont vendu du rêve*, s'amuse Fakhreddine Amara, le directeur de l'activité Data Intelligence de la société de services Keyrus. *J'ai vu des plans qui transformaient le système d'information d'une entreprise en architecture Big Data en quelques mois. Forcément, quand ces espoirs sont confrontés à la réalité, c'est compliqué.* »

Au-delà des espoirs irréalistes parfois placés dans la 'révolution' Big Data, plusieurs éléments se conjuguent pour expliquer pourquoi nombre de POC (Proof-of-concept, soit des prototypes) n'ont débouché sur rien. « *C'est le cas de la majorité d'entre eux* », explique même Emmanuel Manceau, manager chez Quantmetry. D'abord, il y a la question centrale de la donnée et de sa gouvernance, y compris pendant la phase de production. Or, les projets Big Data passent souvent par une mise en commun de données issues de différents métiers de l'entreprise. « *Il faut donc créer une culture de la transversalité* », résume Emmanuel Manceau. Sans oublier, évidemment, la question de la mise en qualité des données. Un chantier central, mais pour lequel les entreprises dédient souvent peu de moyens.

« Industrialiser ce qu'on ne maîtrise pas »

Quantmetry préconise par exemple de responsabiliser les métiers producteurs de données sur ce sujet, quitte à leur verser des subventions pour veiller à la mise en qualité. « *Très souvent, on observe une absence de règles claires associées à l'ingestion des données dans le datalake, tant en termes de qualité des données que d'organisation*, ajoute Matthieu Blanc, du cabinet de conseil Xebia. *Ce qui crée un problème d'organisation du datalake. Or, il est très difficile d'industrialiser ce qu'on ne maîtrise pas.* » Un point bloquant qu'on observe d'ailleurs partout dans le monde, assure Xebia. « *Quand on a assisté à la vague de création de datalakes Hadoop, on ne s'est pas assez posé la question de la qualité des données*, abonde Fakhreddine Amara. *Qui, plus est, dans l'écosystème Hadoop, il n'existe pas de solution de gouvernance des données.* » Une lacune sur laquelle convergent nombre d'interlocuteurs. Xebia a d'ailleurs incubé un éditeur baptisé Amalthea qui est en passe de lancer une solution dédiée à la gouvernance des données dans les environnements Hadoop. Sa solution, actuellement en bêta, doit être finalisée mi-2017 ; elle est déjà en test chez des clients de Xebia, dans l'énergie et la finance.

Car si les entreprises bafouillent, c'est aussi qu'elles sont confrontées à un environnement technologique nouveau… et parfois déroutant. Comme le disait dans nos colonnes récemment un

des cadres de la DSI d'AG2R La Mondiale : « *La courbe d'apprentissage de ces technologies n'est pas évidente, d'autant que ces dernières sont en évolution permanente. Mais c'est aussi passionnant et cela permet de dynamiser la DSI.* » Pour le cabinet d'études Gartner, 70 % des déploiements Hadoop échouent à atteindre les économies projetées ou à délivrer les bénéfices attendus.

« Comprendre comment ce truc fonctionne »

« *Le nombre de clients qui ont aujourd'hui réussi à apprivoiser Hadoop est probablement inférieur à 20, peut-être même à 10*, expliquait récemment Bob Muglia, l'ancien dirigeant de Microsoft devenu patron de Snowflake Computing, éditeur qui propose un datawarehouse dans le Cloud. *C'est juste dingue quand on pense au temps depuis lequel cette technologie est sur le marché et à l'énergie que l'industrie en général y a dépensée*. » Même s'il prêche pour sa chapelle, cette saillie de ce vétéran de l'industrie logicielle dans les colonnes de *Datanami* en dit long sur l'ambiance aux Etats-Unis autour du framework Open Source.

Chez nos confrères toujours, Bobby Johnson, un des ingénieurs qui a permis à Facebook d'exploiter efficacement le framework, reconnaît que les attentes entourant Hadoop sont probablement trop grandes. Que la technologie est bien adaptée pour fournir un entrepôt de stockage à bon marché ou pour faire tourner des tâches d'ETL en mode batch. Mais qu'elle l'est beaucoup moins quant il s'agit de s'attaquer à des applications interactives, destinées aux utilisateurs finaux, en particulier en raison de ses performances décevantes. « *Vous devez réellement comprendre comment ce truc fonctionne pour obtenir ce que vous souhaitez* », ajoute Bobby Johnson. Le créateur de Kafka, Jay Kreps, qui a été responsable d'un grand cluster Hadoop chez LinkedIn, ne dit pas autre chose : « *construire sur cette pile logicielle est réellement très complexe. Je pense qu'il s'agit davantage d'un problème technologique que quoi que ce soit d'autre.* »

Comprendre comment Hadoop fonctionne, en particulier lorsqu'il s'agit de le confronter aux impératifs de la production, suppose d'aligner des équipes importantes. « *Souvent des équipes plus nombreuses que celles qu'on connaissait dans la BI classique*, note même Fakhreddine Amara. *Ce qui, in fine, freine les investissements, car les clients ne voient pas se matérialiser les économies promises par l'Open Source.* » Ou quand l'argument n°1 de Hadoop – les économies qu'il amène par rapport aux architectures traditionnelles – se retourne contre lui.

Un Data Scientist n'est pas un ingénieur logiciel

Ces questions techniques sont intimement liées à des aspects RH. Car, fondamentalement, les solutions de l'écosystème Hadoop sont pensées pour des Data Scientist. Mais, quand il s'agit de passer en production, ces profils si prisés ne suffisent plus. « *Il faut alors leur associer un architecte et des ingénieurs data, pour assurer le pilotage des flux de données, mais aussi un directeur de programme, chargé de l'animation des métiers et du pilotage des ressources techniques. Or, c'est peut-être là que réside aujourd'hui la plus grande pénurie de profils qualifiés* », estime Emmanuel Manceau. Le passage en production implique aussi de reprendre le code développé par les Data Scientists pendant la phase de prototypage. « *Les Data Scientists, qui développent les applications du Big Data, ne sont pas des ingénieurs logiciel. Ils ne sont pas forcément formés aux pratiques des tests automatisés ou à la réutilisation du code* », remarquait, lors d'une conférence sur le salon Big Data, Damien Bigot, responsable data

et outils marketing de Voyages-SNCF. Chez le voyagiste, la phase de passage en production de l'application de recommandation de destinations, qui encaisse une volumétrie de 10 millions de recommandations par jour, a duré près d'un an.

Enfin, les errements des initiatives Big Data masquent aussi des enjeux de pouvoir. Entre les métiers et la DSI, les premiers ayant parfois pris l'initiative et tentant, dans un second temps, d'embarquer l'IT. Dans d'autres cas, c'est l'informatique qui a pris les devants et se heurte à un manque d'appétence (ou de temps) des métiers. Dans certaines grandes entreprises, plusieurs projets ont même ainsi démarré en parallèle, « *tant le sujet était martelé par le marché* », observe Fakhreddine Amara, pour qui ces débuts un peu débridés expliquent aussi le fort ratio d'échecs une fois que les organisations commencent à rationaliser leurs initiatives.

Quand le GDPR percute les datalakes Hadoop

Ces initiatives sont aussi, parfois, rattrapées par le service juridique, qui s'interroge sur l'exploitation des données collectées et agrégées. D'autant plus facilement que le règlement européen sur la protection des données (GDPR), qui entrera en vigueur en mai 2018, pousse les organisations à balayer les traitements de données personnelles qu'elles opèrent. Dans nos colonnes, récemment, Florian Douetteau, le co-fondateur de Dataiku, comparait l'effet du GDPR à des réglementations nord-américaines comme Sarbanes Oxley, poussant les entreprises à se pencher sur l'auditabilité de leurs processus analytiques. « *Par exemple, si on effectue un profilage, il faudra préciser (avec le GDPR, NDLR) les données qui sont utilisées, y compris celles de tiers. Dans la plupart des cas, cela signifie une reconstruction de bout en bout des processus analytiques* », expliquait-il. « *Ces contraintes-là et la sécurité ralentissent les passages en production. D'autant qu'il n'existe pas de solution clef en main pour y faire face* », abonde Fakhreddine Amara.

Toutes ces questions sont toutefois le signe d'une maturité grandissante du marché face aux traitements Big Data. Comme en témoigne d'ailleurs l'offre d'un acteur comme Keyrus. « *On propose aujourd'hui d'anticiper sur les solutions et sur la manière dont on va les déployer*, explique Fakhreddine Amara. *Par exemple via une offre Cloud facilitant le passage à l'échelle ou via une étude d'industrialisation menée au milieu du POC.* »

IA : les attentes irréalistes des dirigeants

La rapidité de l'évolution de l'écosystème technologique ne manquera toutefois pas de confronter les entreprises à de nouvelles difficultés. En particulier si ces dernières placent des espoirs démesurés dans les algorithmes de Machine Learning ou d'intelligence artificielle, la couche d'analyse qui vient très rapidement une fois l'infrastructure Big Data en place. « *D'abord, l'adoption de ces technologies masque un sujet RH*, dit Emmanuel Manceau de Quantmetry. *Il faut commencer par rendre les modèles intelligibles par les utilisateurs, en travaillant sur l'explication des variables. C'est ce qui permettra de transformer des alertes émanant du système en consignes. Le modèle de la boîte noire ne fonctionne pas.* » Et d'alerter aussi sur le décalage entre les craintes du terrain concernant l'IA ou le Machine Learning et les attentes des comités de direction.

Des attentes pour tout dire parfois un brin déraisonnables. Car les algorithmes ne constituent pas

la solution miracle à tous les problèmes complexes auxquels font face les entreprises. « *Une fois le modèle de Machine Learning en place, le travail ne s'arrête pas là* », avertit Emmanuel Manceau, tordant le cou à une croyance assez répandue parmi les dirigeants. Car, la performance des modèles va se dégrader dans le temps. Dans la détection de fraudes, mais aussi dans la maintenance préventive. « *Un modèle modifie le comportement qu'il observe. Conséquence : la base d'apprentissage se dégrade et, avec elle, la performance du modèle* », précise le manager de Quantmetry. Pour qui, c'est toute une stratégie qu'il faut mettre en place et faire vivre dans la durée : AB Testing pour mesurer l'efficacité des modèles, ressaisie des décisions du modèle par les opérateurs pour maintenir la base d'apprentissage, assemblage de différents modèles travaillant sur différentes variables, ré-entrainement régulier des algorithmes… Des contraintes que connaissent déjà bien les sociétés qui se sont frottées très tôt à ces technologies, Criteo effectuant par exemple trois ré-entraînements par jour de ses algorithmes, en mobilisant les données les plus récentes.

L'approche mathématique fait des miracles… parfois

Si le chemin s'annonce donc encore pavé d'embûches, les bénéfices sont bien réels dans certains cas. Fakhreddine Amara cite par exemple le cas de la maintenance préventive des portes de trains, très impactées par les dévers des quais. « *Dans ce scénario, l'approche mathématique fait des miracles* », assure-t-il. Ce serait aussi le cas avec le placement des taxis dans une ville. Un scénario où l'algorithme parvient aujourd'hui à 85 % de confiance (autrement dit, des indications justes dans 85 % des cas, si on compare les résultats obtenus sur les données de test aux données réelles). « *Aujourd'hui, cette compagnie de taxis se fie aux prévisions de demandes du modèle*, explique le dirigeant de Keyrus. *Mais, dans d'autres cas, le Machine Learning n'amène rien, car le résultat n'est pas garanti. Par exemple, si on tente de prédire les réservations de billets d'avion en fonction de la météo. Une forme de corrélation existe entre ces variables, mais le niveau de confiance est trop insuffisant pour espérer mettre une application en production.* »

A lire aussi :

IA : en 2020, un marché qui pèsera autant que le CRM

Florian Douetteau, Dataiku : « Le GDPR va remodeler les applications Big Data »

Pour Air France, le Big Data est un atout maître dans la relation client