LeMagIT
L'informatique pour et par les pros
TechTarget
E-handbook
Contenu PRO+

Le catalogue de données, un pilier de la gouvernance

Dans ce guide

Introduction.

Ils sont l'un des moteurs de la gouvernance des données. Parce qu'ils donnent la possibilité aux entreprises de créer un instantané de l'ensemble de leur patrimoine informationnel, les catalogues de données sont aujourd'hui au cœur des stratégies de transformation des entreprises.

Ces outils garantissent certes une vision générale des données, quelles que soient les sources, mais les exposent, par le biais de tris intelligents, à des populations métiers cibles. Là est l'intérêt : les métiers ont à disposition de bonnes données, en libre-service et n'ont aucun frein pour les utiliser.

Si cela a déjà de quoi convaincre, il est un cas d'usage qui monte : celui de la conformité. En remontant les métadonnées et en les classant, ces catalogues assurent un suivi de leur pédigrée et de leur conformité. Un bienfait lorsque l'on est contraint de suivre l'évolution des réglementations et de les mettre en pratique – le GDPR en est un exemple. Mais cette traçabilité permet aussi de

Dans ce guide

garantir un niveau de qualité adéquate lorsqu'elles sont utilisées par les métiers. De quoi alors former un cercle vertueux dans l'entreprise.

Ce Guide Essentiel a pour vocation d'accompagner les entreprises dans la compréhension de ce concept. Une mise à l'étrier pour entamer une stratégie de gouvernance des données et mieux servir les métiers.

Dans ce guide

Le catalogue de données : à la recherche de l'ordre perdu

Cyrille Chausson, rédacteur en chef LeMagIT

En proposant de cartographier et de classer les données d'un SI, les catalogues de données redonnent la parole au patrimoine informationnel de l'entreprise et favorisent l'usage des données auprès des métiers.

Depuis plusieurs mois, le monde du data-driven en parle, les entreprises qui ont fait le choix de piloter leurs activités en analysant leurs données y voient une pilule miracle : les catalogues de données.
Ce marché, inclus plus spécifiquement dans celui de la gestion de données, correspond à un segment naissant pour le cabinet d'analyste Gartner. Les briques technologiques et les acteurs se mettent en place au fur et à mesure que les entreprises considèrent cet outillage comme une clé de leur stratégie liée à la donnée. Mais sans ambages, Gartner les considère déjà comme une technologie d'avenir, un segment dont la croissance sera fulgurante.

Il faut dire que le catalogue de données a quelques cartes à jouer en matière de gouvernance des données, la pierre angulaire de toutes stratégies dite « data-driven ». Son principe consiste schématiquement à exploiter les métadonnées

Dans ce guide

pour redonner une classification concrète de l'information et de la rendre ainsi « consommable » par les métiers. Là est l'ambition ; ce que Zeenea, l'un des pure-players français de ce segment considère comme étant de la « démocratisation des données » au sens large.

Le catalogue de données est en effet là pour livrer un instantané des données, de localiser leur emplacement, d'identifier les jeux les plus exploitables pour une population de métiers donnée et de montrer leur pédigrée en détaillant leur traçabilité (data lineage).

L'étape d'après consiste à les trier, les étiqueter, les classer. Là le Machine Learning apparait comme un outil essentiel pour nombre d'acteurs qui automatisent « intelligemment » cette étape. Des métadonnées et du contenu même des données sont extraites des informations qui livrent une identité à une information, avec un scoring (cela est répandu chez nombre d'éditeurs) qui évalue la pertinence de cette reconnaissance.

Techniquement, le catalogue de données scanne l'ensemble des systèmes de l'entreprise pour y collecter automatiquement les métadonnées.

De là est également évalué le parcours de cette donnée – sa traçabilité. L'intérêt est ensuite d'indexer ces métadonnées afin d'en faciliter la recherche et la consultation par les métiers ou les analystes. « L'intérêt consiste donc à aller scanner au plus large le SI de l'entreprise », explique ainsi Edouard Guérin, consultant avant-vente

Dans ce guide

Big Data et gouvernance des données chez Informatica, et d'y inclure chaque parcelle de l'IT afin de cartographier l'ensemble des sources.

Alation, Informatica, Zeenea, Waterline Data proposent tous une interface proche de celle de Google pour justement faciliter l'accès à ces informations par des non techniciens IT.
« Cela pose un canal d'échange commun entre l'IT et les métiers et permet d'avoir une vision précise de ce qui produit de la valeur pour l'entreprise » - Stéphane Jotic, CEO Zeenea (pure-player français du catalogue de données).

Cette interface à la Google constitue en fait une porte d'entrée très large pour les métiers. Mais il n'agit que comme un premier filtre pour un service encore plus ambitieux : le portail qui permet de l'utiliser en libre-service par certaines populations d'utilisateurs. L'objectif est de « proposer un service aux métiers afin de les rendre autonomes et d'agir avec des fonctions préparées par ces mêmes métiers. Sans catalogue qui a précédemment identifié les données, difficile de mettre en place un tel service », ajoute encore Edouard Guérin.

Sur ce point, le marché se divise encore, certains éditeurs ayant intégré cette fonction dans leur offre. Chez d'autres il s'agit d'un second produit.

Aussi automatisé soit-il, le catalogue de données passe ensuite la main aux experts « humains ». Ils ont la possibilité, de façon collaborative, de valider les informations ou de suggérer des corrections. Celles-ci sont en revanche

Dans ce guide

effectuées dans un outil tiers (par exemple Trifacta), la qualité des données n'étant pas une fonction du catalogue de données.

Du data lake à la recontextualisation

Si l'on comprend ici l'intérêt de la technologie, la question à se poser est pourquoi ces catalogues ont le vent en poupe aujourd'hui. Comme tout catalogue, celui de la donnée a pour ambition d'organiser le patrimoine informationnel de l'entreprise en classant l'ensemble de ses données. Trier méthodiquement en somme, alors que l'ère du Big Data, il y a 10 ans avait justement poussé les entreprises à placer leurs données dans de vastes systèmes, toutes au même niveau, quel que soit leur format, sans hiérarchisation, sans tri. Un énième système – un data lake Hadoop par exemple – était donc venu s'ajouter à ceux en place, accentuant un peu plus la dilution de l'information et la répartition massive des données, en silo.

« Plus de 72 % des entreprises n'ont pas une culture de la donnée, même si elles ont investi massivement dans le Big Data et l'AI », assure Satyen Sangani, le Pdg d'Alation, éditeur américain d'un catalogue de données, citant une étude du cabinet New Vantage Partners.

Sans cette culture infusée à l'ensemble de l'entreprise, l'édifice se fissure. « Les investissements ont certes été massifs, mais les résultats limités », ajoute encore le spécialiste. « D'où la nécessité de re-classer les données », pour les

Dans ce guide

porter au plus près des usages métiers. Il s'agit là de « révéler les données telles qu'elles existent », illustre-t-il.

Ces investissements ont été par exemple consentis dans un outillage stratégique, comme dans des solutions de Business Intelligence (BI). Mais ils n'ont pas atteint leur plein potentiel, freinés justement par ce manque de culture. Et ces entreprises-là sont les premières aux abois. « Nous ciblons ainsi les entreprises qui ont un pied dans les outils de BI », reconnait encore Satyen Sangani. Et le cercle est vertueux, car en révélant ces données, « le catalogue contribue à apporter cette culture de la donnée ».

Cet argument déceptif est également repris par Edouard Guérin (Informatica). « Beaucoup de promesses n'ont pas été tenues par le passé (il y a 10 ans environ, quand le Big Data a pris forme, NDLR). Le data lineage a par exemple été placé dans un seul contexte décisionnel, alors que le catalogue de données ne doit pas être contraint à cet environnement, mais doit proposer une vision globale. »
Seulement voilà : « A l'époque, la puissance de calcul ne permettait pas d'avoir cette vision globale », lance-t-il.

Pour Zeenea, la montée en puissance des catalogues de données est aussi venue avec la prise en compte des données non structurées – et donc des environnements Hadoop. De vastes lacs (parfois marécages diront certains) où toutes les données, quels que soient leurs formats, sont stockées au même niveau sans classification.

Dans ce guide

« Le catalogue de données crée ainsi un canal d'échange commun entre l'IT et les métiers », résume encore un responsable de la société et permet ainsi d'avoir une vision précise de ce qui produit de la valeur (dans des environnements de plus en plus pilotés par la donnée, NDLR) ».

Gouvernance des données et RGPD

Evidemment, le catalogue de données trouve naturellement sa place dans les politiques de gouvernance de données. « L'intérêt [de celui-ci] tient au fait que la donnée soit devenue un élément clé de l'entreprise, explique Edouard Guérin d'Informatica.
« La valeur des données ne peut s'exprimer qu'à condition de bien les connaître et de savoir où elles se trouvent. » Les cartographier ainsi que leurs sources, mais aussi suivre leur traçabilité et analyser leurs impacts à l'échelle du SI, sont les autres piliers de ce catalogue. On peut par exemple comprendre l'impact qu'a une montée de version d'un logiciel sur une autre application ou un autre jeu de données, illustre-t-il.

Mais l'autre moteur est certainement la réglementation. En proposant des capacités de traçabilité, le catalogue de données devient un allié clé pour toutes formes d'audits.

« Les contraintes réglementaires, comme le RGPD et la norme IFRS 17 pour le monde de l'assurance représentent aujourd'hui un moteur pour le catalogue de

Dans ce guide

données », commente le spécialiste. Avec des SI distribués entre plusieurs pays, les données sont en plusieurs langues et répondent à des normes locales. « Le catalogue des données scanne tout le SI, interne et externe et crée une radiographie complète des systèmes, du mainframe aux objets connectés. Dans un contexte RGPD, il identifie les données à caractère personnel d'une part et aide à constituer un registre des traitements. »

Dans ce guide

Les étapes pour construire un catalogue de données

Anne Marie Smith, Ph.D., Alabama Yankee Systems, LLC

Un catalogue de données est le garant des métadonnées et des données de l'entreprise. Mais sa conception demande un peu de méthode. Cet article vous accompagne dans sa mise en œuvre.

Si agréger les données est une première étape, les rendre accessibles à la bonne population d'utilisateurs en est une autre, tout aussi importante. C'est là qu'entre en jeu le catalogue de données. Sa création devient aujourd'hui un enjeu majeur pour parvenir à utiliser la donnée comme un vrai actif de l'entreprise. Mais attention, elle est aussi un processus collaboratif. Les entreprises ne doivent pas entreprendre un tel projet sans la contribution de leurs partenaires commerciaux, ni les départements métiers.

Un catalogue de données est une application de référence qui permet aux utilisateurs métiers et IT d'explorer les sources de données, de comprendre leur contenu via des métadonnées, de connecter ces données à la source et d'y accéder en toute autonomie – en libre-service. Un catalogue de données explore donc les bases de données et les systèmes de BI. Il fournit également

Dans ce guide

un point de référence unique pour la gestion des métadonnées de l'entreprise, plus rapide et plus efficace que les anciens systèmes.

Les principales étapes de la création d'un catalogue de données sont les suivantes:

- **Concevoir un modèle de données qui servira de base à l'architecture du catalogue.** Un catalogue de données efficace doit correspondre à l'usage des données par les métiers, et ne pas être une simple implémentation technique. Un modèle doit (SAM – Subject Area Model) définir chaque sujet et concepts associés. Il montre aux métiers l'emplacement de leurs données sans référence aux applications, aux fichiers ou aux bases de données. Le catalogue de données doit être construit sur la base de ce SAM.

- **S'appuyer sur les Data Stewards et les responsables IT pour découvrir et accéder aux métadonnées de toutes les bases de données et tous les fichiers.** Les catalogues de données utilisent des métadonnées pour identifier les tables, les fichiers et les bases de données. Pour cela, il effectue une recherche dans les bases de données de la société et charge les métadonnées (et non les données réelles) dans son référentiel. Avant toute création, les sources des métadonnées doivent être identifiées puis enregistrées. Il s'agit d'une étape majeure qui nécessite un solide programme de gouvernance. Les

Dans ce guide

Data Stewards sont ici importants pour avoir un aperçu des sources de données à utiliser.

- **Construire un dictionnaire de métadonnées (pas un glossaire métier).** Ce dictionnaire contient la description et le mapping de toutes les tables ou fichiers et de toutes leurs métadonnées. Ce dictionnaire devient la base du catalogue de données. Là encore, les Data Stewards métiers sont essentiels car ils identifient les métadonnées à utiliser dans le catalogue - par source, concept et domaine.
- **Profiler les données pour proposer des statistiques aux utilisateurs.** Ces profils sont des résumés informatifs qui expliquent et aident à comprendre les métadonnées. Par exemple, le profil d'une base de données comprend souvent le nombre de tables, de fichiers et le nombre de lignes.
- **Identifier les relations entre les sources.** Il s'agit là de découvrir les données associées sur plusieurs bases de données. Un analyste peut par exemple avoir besoin d'informations consolidées sur le client. Grâce au catalogue de données, on peut noter que cinq fichiers sur cinq systèmes différents contiennent des données client.
- **Développer une traçabilité des données.** Les outils d'ETL (Extract, Transfer, Load) sont utilisés pour extraire les métadonnées des bases de données sources, les transformer et les nettoyer, puis les charger

Dans ce guide

dans une base de données cible. Cela peut être utile pour rechercher les éventuelles erreurs d'une l'analyse.

- **Structurer le catalogue pour l'humain (en fonction du SAM).** La plupart des fichiers et bases de données sont conçus pour être utilisés par des outils technologiques. Les catalogues de données doivent être conçus tant pour ceux qui consomment les données que pour ceux qui fabriquent les technologies. Autre élément clé : un catalogue de données doit rester consultable depuis un ordinateur, une tablette et des applications mobiles.

Dans ce guide

La traçabilité des données : un turbo pour la gouvernance

David Loshin, Knowledge Integrity Inc.

La gouvernance des données est essentielle pour les entreprises qui souhaitent suivre le cycle de vie des données. Cet article donne quelques conseils sur les points à considérer si on souhaite investir dans la traçabilité des données.

La raison première de la gouvernance des données est le respect et la conformité de la politique en matière de données dans une entreprise.
Ces politiques peuvent couvrir de nombreux objectifs, et reposer sur des directives sur la protection et la validation des données.

Les responsables de la gestion et de la gouvernance des données doivent pour cela solliciter les utilisateurs métier.
Il s'agit de formuler clairement les exigences en matière de qualité des données, de préciser les paramètres et de mesurer la conformité aux politiques de données.

Cependant, le défi est de combler le fossé qui existe entre la définition même de ces politiques de gouvernance des données et leur mise en œuvre. Les

politiques visent à assurer le contrôle de la qualité des données dans l'ensemble des flux de production. Toutefois les responsables qui se voient souvent confier la responsabilité de la gestion de la qualité des données restent sans formation ou sans outils appropriés.

C'est là qu'interviennent les outils de traçabilité de la donnée (data lineage). Cette fonction documente le parcours des données dans l'entreprise et aide à simplifier deux procédures clés de la gouvernance des données : l'analyse des causes et l'analyse d'impact.

Traçabilité et gouvernance des données

Si l'on ne dispose pas d'un moyen pour identifier où sont les erreurs, les responsables des données (que l'on appelle les data steward) auront du mal à identifier et à corriger les problèmes en matière de qualité des données.

Lorsque ces erreurs continuent de se propager, l'entreprise risque d'être confrontée à des rapports et des analyses incohérents – et donc de mauvaises décisions.

Les outils de data lineage (suivi des données) peuvent simplifier le processus d'analyse de causes fondamentales en cartographiant les différents traitements par lesquels les données sont passées.

Dans ce guide

La qualité des données peut être examinée à chaque point du flux de traitement, ce qui permet à l'IT de trouver l'origine des erreurs.

En remontant à cette erreur première, le data steward peut insérer des contrôles à chaque étape pour vérifier si les données étaient conformes aux attentes ou si l'erreur était déjà présente.

L'étape qui indique que les données étaient conformes à l'entrée, mais défectueuses à la sortie, est celle où l'erreur a été introduite. L'administrateur des données peut donc se concentrer sur l'élimination de la cause fondamentale au lieu d'essayer simplement de corriger les mauvaises données.

Tracer l'historique des données peut également aider les data steward à identifier des changements inattendus de format et de structure des données - les environnements actuels sont en effet beaucoup plus dynamiques que dans le passé. Lorsque les sources de données changent, il peut y avoir des conséquences imprévues.

A partir de son origine, le gestionnaire des données peut également retracer les dépendances et déterminer les étapes de traitement impactées par le changement.

Dans ce guide

Ce qu'il faut rechercher dans les outils de traçabilité des données

La collecte manuelle des métadonnées et la documentation du data lineage nécessitent un investissement important en ressources.

Toutefois, elles restent sujettes à l'erreur, surtout dans les entreprises qui s'appuient sur des rapports et des analyses pour la prise de décision.

Il convient alors de rechercher des produits qui permettent de :

- Accéder de manière native à un large éventail de sources de données,
- Regrouper les métadonnées dans un référentiel centralisé,
- Fournir des présentations simplifiées des métadonnées à différents utilisateurs et encourager la collaboration pour aider à la validation des métadonnées,
- Documenter la façon dont les données circulent dans les flux de traitement,
- Fournir une présentation visuelle de la traçabilité des données,

Dans ce guide

- Fournir des API aux développeurs pour interroger les informations de traçabilité,
- Créer un index inversé pour faire correspondre les éléments de données à leurs usages,
- Fournir des modules de recherche pour retracer rapidement le flux de données depuis son point d'origine jusqu'à toutes ses cibles en aval.
- Parcourir les flux de données.

Dans ce guide

Les conseils de Zalando pour rendre votre entreprise « Big Data-driven »

Philippe Ducellier, journaliste LeMagIT

Lors du Salon Big Data Paris, le « Monsieur Données » du e-vendeur d'articles de mode a partagé ses bonnes pratiques pour concevoir, mettre en production et infuser le Machine Learning dans une organisation pour que les modèles soient utilisés (et réutilisés).

Khitij Kumar est vice-président en charge des données chez Zalando, l'un des plus grands magasins en ligne d'Europe pour la mode. « Zalando a commencé comme une simple e-boutique, il y a dix ans », explique-t-il.

Aujourd'hui, le site allemand vend 300.000 articles, de 2 000 marques internationales, pour un chiffre d'affaires de 5,3 milliards d'euros.

Mais il ne vend pas de manière uniforme. Le distributeur 100 % web cherche en effet à localiser et à personnaliser son assortiment le plus possible pour augmenter les ventes et réduire les délais (et les coûts) de livraisons.

« Ce que nous vendons à Paris est différent de ce que nous vendons à Londres ou à Dublin », confirme Khitij Kumar. « Au final, Zalando est un facilitateur.

Dans ce guide

Quelqu'un produit des choses. Quelqu'un consomme ces choses. Nous mettons les deux en relation. Et pour y arriver, tout que nous faisons s'appuie depuis le début sur la donnée - de la BI à l'Intelligence Artificielle en passant par la Data Science ».

Zalando est donc bien un distributeur, mais il se présente également - voire surtout - comme une entreprise technologique.

« Nous sommes probablement aujourd'hui l'une des sociétés les plus avancées au monde dans le Big Data », affirme Khitij Kumar. « Quelque soit l'outil [BI ou Data Science] que vous pouvez nommer, il y a de fortes chances que nous l'ayons déjà regardé, et même probablement utilisé [...] Nous faisons beaucoup de Data Warehouse avec des bases de données traditionnelles. Nous avons aussi un Data Lake dans le cloud. [...] Nous sommes dans le cloud - pas juste un cloud, plusieurs cloud - nous sommes très avant-gardistes et très scalables. Et nous embauchons dans le domaine de la technologie ».

Pour preuve, à date, Zalando compte plus de 2000 employés dans l'IT (sur 15.000 employés), la plupart basés à Berlin.

Dans ce guide

Mettre les données au cœur des processus et impliquer des leaders

L'essor Pourquoi toute cette expertise technique ? Parce que Zalando veut que toutes ses actions s'appuient sur l'analyse des données qu'il collecte chaque jour.

« Toute l'expérience que vous avez sur le site Web découle du Machine Learning (ML) », explique le Monsieur Données de Zalando sur la scène de Big Data Paris.

« Par exemple, nous faisons évidemment le "si vous achetez ça ou si vous regardez ça, alors ceci pourrait vous intéresser" ».

Le matching peut se faire sur différents paramètres : « un style qui correspond au genre que vous regardez, ou une fourchette de prix qui correspond à ce que vous avez déjà acheté, etc. ».

Les algorithmes sont également appliqués pour le cross-selling à la volée : « si je trouve une veste, j'aurai peut-être besoin d'un pantalon, de chaussettes, de chaussures. J'ai peut-être aussi besoin d'une cravate ».

Bref, la donnée et l'apprentissage statistique sont au cœur de nombreux processus de Zalando, et pas seulement du site. Khitij Kumar souligne en

Dans ce guide

préambule que cette diffusion du Big Data dans son entreprise vient d'une volonté des dirigeants.

« Le Big Data a besoin de quelqu'un qui soit prêt à prendre le taureau par les cornes [...] Il faut avoir une attitude d'entrepreneur », prévient-il.

Chez Zalando, la gestion des stocks et de la chaîne d'approvisionnement, le marketing et les ventes, ou encore la publicité sont tous infusés avec du Big Data et du Machine Learning.

« Lorsque nous achetons des stocks pour l'année prochaine, nous voulons nous assurer que nous achetons les bonnes choses et que nous vous livrerons ce que vous voulez au moment où vous le voudrez. Dans le marketing, lorsque vous allez sur Facebook ou sur tout autre site, nous voulons nous assurer que nous vous affichons le bon message sur ce que vous voulez vraiment. En matière commerciale, nous voulons vous proposer les meilleurs rabais possibles [qui permettent de conclure une vente] etc. »

Et tout cela avec de l'analytique Big Data pour ne pas simplement proposer un peu plus vite ce que la personne cherche, mais pour lui proposer ce qu'elle *pourrait* chercher, « même quand elle-même n'a pas encore pris sa décision ».

Machine Learning et Big Data s'articulent donc intimement avec de véritables problématiques métiers. Ce qui ne va pas sans une certaine organisation (lire ci-après).

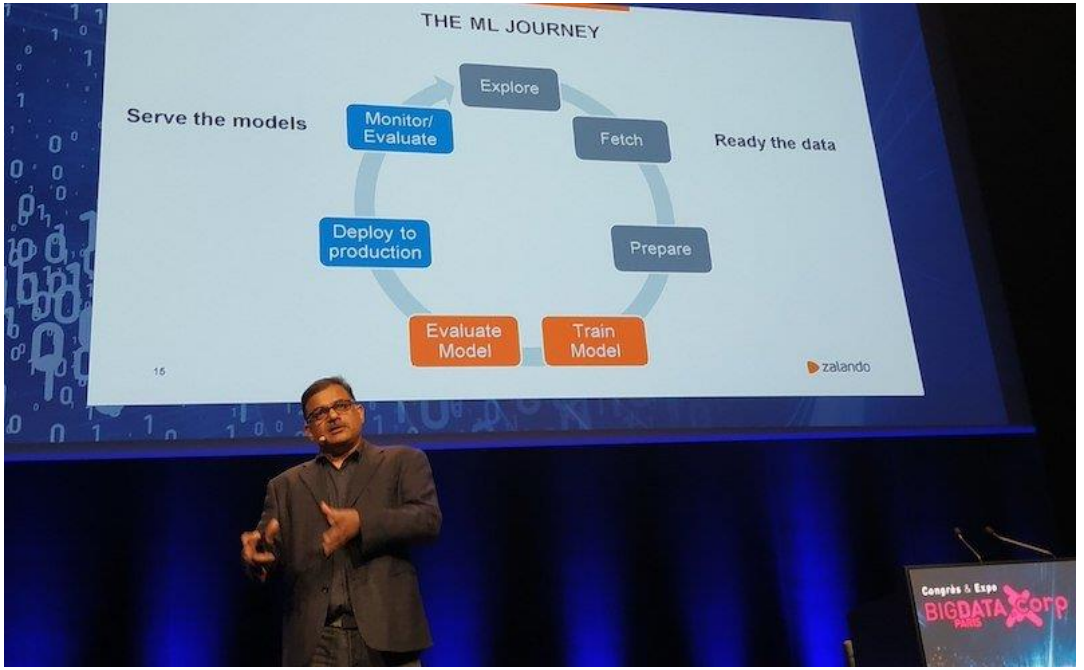
Dans ce guide

« Nous autorisons par exemple une politique de retour de 100 jours. Vous pouvez acheter ce que vous voulez sur le site et nous le retourner. Typiquement, les gens essaient quatre ou cinq articles pour n'en garder qu'un seul. Mais si l'on arrive à vous proposer exactement ce que vous voulez en une seule fois, cela économise beaucoup de temps [et d'argent] ».

Étapes d'un projet de Machine Learning et bonnes pratiques

Khitij Kumar conseille que tout bon projet ML commence par l'exploration des données, leur compréhension, leur extraction et leur préparation.
« Vous ne pouvez pas faire de ML si vous n'avez pas nettoyé les données et retiré celles qui sont erronées ».

Dans ce guide



Etapes d'un projet Big Data et Machine Learning, Khitij Kumar à Big Data Paris 2019

La deuxième étape consiste - évidemment - à créer le modèle et à le former. Une fois l'entrainement du modèle terminé, la troisième étape consiste à le vérifier et à l'évaluer pour voir s'il fonctionne réellement et comme prévu.

Pour faire tout cela (et les étapes suivantes), Zalando utilise principalement des outils open source. « Nous construisons beaucoup nos outils nous-mêmes,

parce que dans de nombreux cas nous faisons des choses qui nous sont très spécifiques ».

Néanmoins, Zalando n'est pas opposé par principe à l'utilisation de produits propriétaires. « Mais nous regardons toujours en premier lieu l'open source », insiste le responsable.

Le conseil de Khitij Kumar sur ce point est de bien regarder ce que vous faites en interne et de voir s'il n'y a pas déjà un outil, un modèle ou une compétence que vous pourriez utiliser. Si ce n'est pas le cas « travaillez avec chacun des éditeurs pour vous assurer qu'ils sont bien capables de vous fournir tout ce que vous n'avez pas pu obtenir par d'autres moyens ».

Suivre les modèles en production

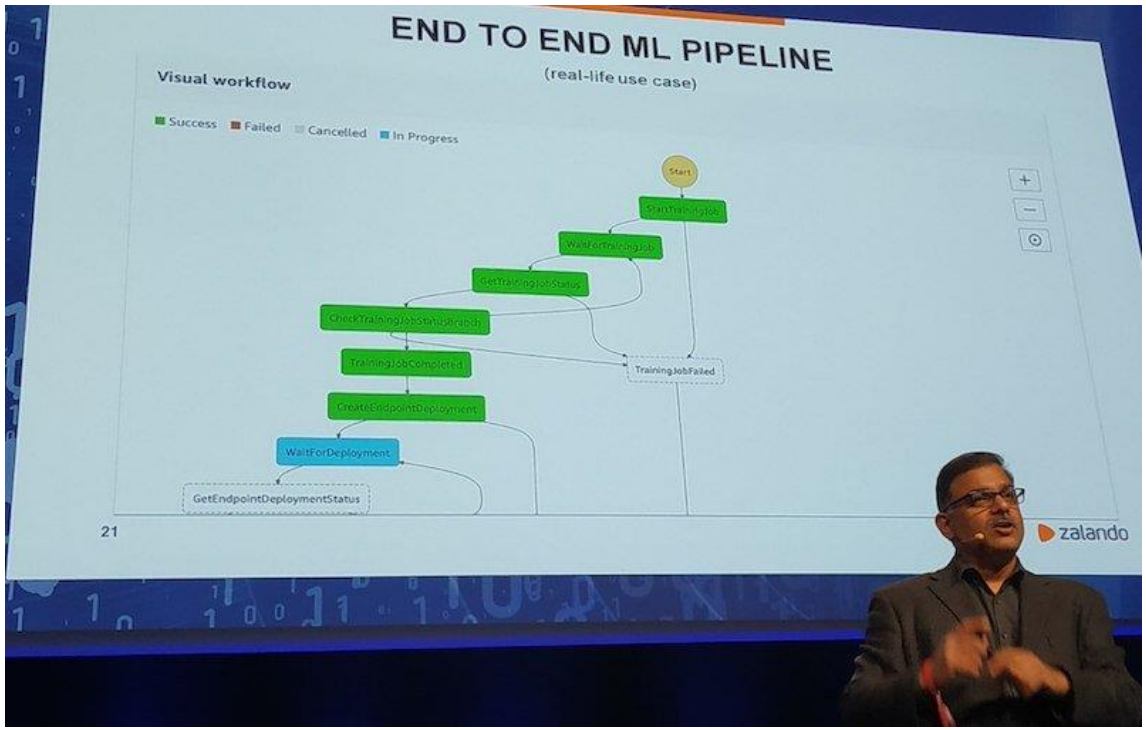
Vient ensuite la mise en production. « Une fois qu'il tourne de manière opérationnelle, vous devez surveiller le modèle et voir s'il se comporte comme prévu », préconise Khitij Kumar.

« Vous devrez être en mesure de regarder, de superviser et d'obtenir des retours sur la façon dont un modèle déployé fonctionne, comme son utilisation du CPU, son utilisation de la mémoire et son utilisation du disque sur une certaine période de temps ».

Dans ce guide

Pour y arriver, le responsable préconise de réaliser un pipeline « très classique » (sic) mais aussi complet du cycle du Big Data.

Ce workflow devra représenter où sont les données, comment les ingérer, l'entrainement des modèles, leurs déploiements et prévoir de regarder ce qui se passe en production pour, si besoin, les corriger et itérer. « C'est une histoire sans fin », plaisante Khitij Kumar.



Khitij Kumar à Big Data Paris 2019

Dans ce guide

Modèles partagés et catalogues de données

Le Pour réellement infuser toute sa structure au Big Data, Zalando a aussi - et surtout - créé une fonction centrale de support du Machine Learning et du Big Data (celle que dirige Khitij Kumar). « Elle couvre tout : Machine Learning, BI, Data Lake, Spark, algorithmes, gouvernance ».

Le but est d'aider les volontaires internes à se lancer. Chaque équipe peut avoir ses propres responsables et experts en ML, mais ils travaillent avec le support de l'organisation centrale.

Car chez Zalando, un modèle doit être disponible et réutilisable par tous. « Par exemple, lorsque vous cherchez une chemise sur notre site, nos équipes de recommandation ont travaillé sur le type de vêtements que nous devrions vous montrer ».

A titre d'exemple concret, Khitij Kumar s'imagine en voyage professionnel et qu'il doit acheter une veste dans un pays étranger pour une réunion. « Si le site me montre une veste à 59,99 euros et pas à 200 euros c'est que l'algorithme a peut-être été en mesure de voir que j'étais en déplacement et que je ne voulais pas acheter des choses que je ne garderais peut être pas. Et donc, il me montre une veste bon marché mais élégante ».

Dans ce guide

Une équipe a pu créer un tel moteur de recommandations avec un paramètre de qualification géographique. « [Mais] il y a d'autres personnes au sein de l'entreprise qui peuvent vouloir l'utiliser : pourquoi ne pas le leur donner et les laisser faire ? [...] Il faut donc que ce modèle soit rendu disponible pour tous nos employés. Il faut aussi s'assurer que les gens le sachent et sachent comment l'utiliser en toute sécurité ». D'où l'intérêt d'une fonction centrale dédiée au Big Data pour rappeler ces points et aider au partage.

Zalando ne veut pas s'arrêter aux modèles de Machine Learning. « Si vous utilisez MicroStrategy (NDR : dont Zalando est client pour la BI) ou Tableau et que vous créez des APIs, pourquoi ne pas le partager avec tous les autres ? »

Pour y arriver, Zalando dispose de catalogues de données mais, ajoute Khitij Kumar, « un catalogue de données qui ne liste pas seulement des données stockées dans des tables. Ajoutons-y les modèles, les API, et bien d'autres choses. Bien sûr, il faut faire cela en ayant à l'esprit la sûreté et la sécurité. Il faut donc aussi être en mesure d'avoir une gouvernance sécurisée autour de tout cela »... et donc, à nouveau, une structure dédiée centralisée.

Ce qui ne veut pas dire que tout soit centralisé, bien au contraire. « Certains de nos spécialistes ML les plus pointus ne font pas partie de mon équipe. Ils sont dans les unités métiers, qui utilisent certains des outils que nous mettons à leur disposition ».

Dans ce guide

Dernier avantage, en diffusant les bonnes pratiques et la réutilisation des modèles Big Data, une telle organisation accélère la conception de nouvelles applications opérationnelles à base de Machine Learning. « Les entreprises doivent avoir bien conscience de la nécessité d'une [telle] fonction centrale », conclut Khitij Kumar.

Dans ce guide

Accéder à plus de contenu exclusif PRO+

Vous avez accès à cet e-Handbook en tant que membre via notre offre PRO+ : une collection de publications gratuites et offres spéciales rassemblées pour vous par nos partenaires et sur tout notre réseau de sites internet.

L'offre PRO+ est gratuite et réservée aux membres du réseau de sites internet TechTarget.

Profitez de tous les avantages liés à votre abonnement sur: http://www.lemagit.fr/eproducts

Images; Fotolia

Dans ce guide

Cyrille Chausson | *Rédacteur en Chef*
TechTarget
22 rue Léon Jouhaux, 75010 Paris
www.techtarget.com