DATA ET IA : CLÉS DU TRAITEMENT AUTOMATIQUE DES DONNÉES HÉTÉROGÈNES

ANNABELLE TUGENDHAT (X02)
directrice de la data & IA Factory, La Poste

L'intelligence artificielle rend de nombreux services pour le traitement des données massives et hétérogènes, comme on en trouve dans les organisations de soutien aux entreprises. Encore faut-il avoir une idée précise de ce qu'on cherche à faire et organiser méthodiquement la conduite du projet.

Quel que soit le secteur d'activité, les entreprises traitent un grand nombre de documents en *back office* (service d'appui) : les factures, les commandes, les CV, les réglementations par exemple, mais également briefs clients, appels d'offres, etc. Non seulement la volumétrie de ces documents est croissante (et alignée avec la croissance exponentielle des données échangées, en 2025 il est attendu un volume de 181 zettabytes), mais leur forme est toujours plus hétérogène et le temps pour les traiter est toujours plus réduit. Comment extraire de la valeur de ces documents pour en tirer un fond homogène, analysable, exploitable et disponible, avec un accès unifié, simple et intuitif ?

Des questions préalables

À l'échelle d'un projet d'automatisation du traitement des documents, il se pose plusieurs types de questions préalables : quelles interfaces pour récupérer, consolider, étiqueter la donnée, quelles interfaces et quels usages à destination à la fois des équipes *back office*, des décideurs et de tout autre utilisateur final qui pourrait tirer de la valeur de ces données ? Quels choix techniques pour traiter les documents, les stocker, les trier ? L'objectif de cet article n'est bien sûr pas de répondre à toutes ces questions qui sont dépendantes de l'usage et de la donnée, mais de donner des clés sur un déroulement de projet et d'identifier les acteurs et les décisions à prendre.

Prenons un exemple simple et concret : une entreprise cherche à répondre à des appels d'offres. Comment fait-elle pour récupérer la liste des appels d'offres ouverts, retirer les dossiers de consultation, lire les dossiers, préparer une vue consolidée pour décider de la réponse ou non à ces appels d'offres, préparer les documents administratifs communs, préparer les éléments de réponse, disposer d'une base commerciale (et alimenter le CRM – *customer relationship management* – de l'entreprise) et proposer l'accès à cette donnée en mode self-service pour les différentes directions de l'entreprise, afin d'optimiser le processus par la suite ?

Deux étapes de traitement

La première partie consiste bien à consolider les données de différentes sources : agréger et lister les appels d'offres pertinents à partir des différentes sources, récupérer les dossiers de consultation et analyser les dossiers selon des critères propres à l'entreprise. Il faut également les stocker et les organiser, les étiqueter, les indexer, →


© tippapatt

→ leur attribuer des règles (sécurité, confidentialité). On peut également avoir besoin de faire intervenir des algorithmes de *machine learning* (apprentissage automatique), de l'OCR (reconnaissance optique de caractères) ou de la RPA (automatisation robotisée des processus) pour traiter les documents, en extraire le contenu et le structurer.

Une fois la donnée accessible et consolidée, il faut pouvoir l'exploiter. La structurer est une étape indispensable afin que l'entreprise puisse prendre ou non la décision de répondre à l'appel d'offres, en fonction de critères qui lui sont propres, analyser les cahiers des charges pour préparer la meilleure réponse, fournir toutes les réponses sous le format attendu par l'entité adjudicatrice. On peut également utiliser cette donnée pour alimenter le CRM de l'entreprise, la rendre accessible à toutes les directions de l'entreprise (par exemple, pour identifier de nouveaux axes stratégiques de développement). Cette deuxième partie d'exploitation consiste à créer les interfaces et les flux pour tirer de la valeur de la donnée que nous avons ingérée.

Nous allons utiliser cet exemple tout au long de l'article pour décrire les constructions des deux interfaces, collecte et consolidation, et recherche et décision.

L'interface de collecte et de consolidation

La première question qui se pose avant de se lancer dans un projet data de traitement de documents est celle, comme pour tout autre projet, notamment IT, de son utilité et de sa mise en œuvre : la volumétrie est-elle suffisante pour justifier de s'engager dans un projet data ? Existe-t-il des solutions sur le marché qui permettent de couvrir tout ou partie du projet (la question du *make or buy*, faire ou acheter) ? Est-il possible d'externaliser une partie de la prestation ? Dans notre exemple, nous pouvons déjà considérer les appels d'offres publics qui constituent une volumétrie d'environ 900 par jour ouvré, soit pas loin de 250 000 appels d'offres publics par an. Si l'on ajoute à ceux-ci les appels d'offres passés par les entreprises privées, en fonction de la segmentation géographique, le multiplicateur est énorme. La volumétrie, la variabilité des documents et les contraintes temporelles (vitesse), ces trois facteurs (les 3 V traditionnels du *big data*) nous placent dans le bon contexte d'un projet data.

Les questions initiales une fois le projet lancé

Une fois que nous avons décidé de traiter ce projet et que nous avons convenu du périmètre que nous allons nous-mêmes traiter (par exemple, il paraît utile d'utiliser des solutions d'agrégation sur le marché plutôt que d'en écrire une), il convient de s'intéresser aux conditions matérielles pour héberger la donnée : où va-t-elle être stockée ? Y a-t-il des contraintes liées à l'organisation de l'entreprise pour le stockage des données (*cloud, on-premise*), doit-on la stocker ou pouvons-nous travailler en collectant et traitant directement la donnée, est-il possible de la traiter *via* un système de base de données standard ou la volumétrie et la variété sont telles que nous allons nous diriger vers des bases NoSQL et du *big data* ? Va-t-on la croiser avec d'autres données (ce qui justifierait de la placer dans un *data lake* par exemple) ? Sommes-nous soumis à des règles de confidentialité, d'anonymisation, de traçabilité (RGPD) ; s'agit-il de données personnelles, *a fortiori* sensibles, est-il nécessaire de prévoir un processus de purge ? Toutes ces questions sont à prendre en amont et sollicitent plusieurs services de l'entreprise : la DSI et son urbanisme pour proposer une solution de stockage, le RSSI chargé de la sécurité de l'informatique, le DPO (délégué à la protection des données) sont en première ligne à ce stade, en accompagnement de l'équipe projet et de l'ensemble des équipes IT chargées de la réalisation et de l'exploitation.

Les canaux d'alimentation

Ensuite, il faut définir les modalités de collecte de la donnée, préparer les ouvertures de flux, éventuellement extraire la donnée et la transformer pour l'étiqueter, l'organiser, la structurer et la cataloguer, le tout en respectant la politique de sécurité de l'entreprise et les contraintes citées ci-dessus, y attacher une supervision pour s'assurer de la complétude de la collecte et de la

fraîcheur des données. Deux sous-étapes importantes, qui ne couvrent pas exactement le même périmètre fonctionnel : la mise en place et la supervision des canaux d'alimentation de la donnée (par des flux froids ou chauds) par les équipes informatiques, ainsi que le traitement de la donnée qui est du ressort des équipes data et qui dans notre exemple peut intégrer des algorithmes de *machine learning*, de l'OCR ou de la RPA pour récupérer et structurer la donnée de manière homogène et permettre l'organisation de cette donnée par nature hétérogène.

L'interface de recherche et de décision

Une fois que nous disposons d'une donnée fraîche, structurée, sécurisée et organisée dans l'infrastructure que l'on souhaite, on peut alors l'exposer pour la rendre disponible aux utilisateurs, aux décideurs et à des applications tierces. Une partie importante du projet va être de définir quel usage va être fait de cette donnée. La première pratique va être de fournir une interface de *search* (recherche) pour celle-ci, afin de permettre aux consommateurs de cette donnée de trouver les éléments qui les intéressent. Ensuite, on peut fournir des interfaces de *data visualization* qui vont exposer des indicateurs simples, ou même se brancher sur de l'IA qui aura proposé des recommandations, par exemple. Puis alimenter des applications tierces. Pour reprendre notre exemple des appels d'offres, on peut imaginer une interface qui permette de chercher dans un *dataset* (jeu de données) des appels d'offres existants en fonction de critères définis par l'utilisateur. Des solutions de *dataviz* (visualisation de données) sur le marché permettent de répondre aux besoins d'interfaces, *modulo* le travail réalisé par des *data analysts* pour préparer cette visualisation.

La réponse aux besoins des utilisateurs

Ensuite, on peut – en fonction des appels d'offres choisis par le décideur pour recevoir une réponse – imaginer une interface qui permette de recommander au décideur par la suite les appels d'offres auxquels il est opportun de répondre. Pour que cette recommandation soit de qualité, il faut implémenter des boucles de *feedback* régulières afin d'entraîner le modèle de recommandation. Il est nécessaire d'avoir une coordination solide entre les équipes de *Data Science* et les équipes informatiques. Enfin, on peut apporter des couches supplémentaires d'outillage, une fois la donnée exposée : l'envoyer vers d'autres plateformes pour communiquer avec des personnes extérieures, alimenter un CRM avec les données des appels d'offres, développer une IA qui permettrait d'automatiser les réponses aux appels d'offres, mettre en place une base documentaire. L'équipe projet va porter une bonne compréhension des besoins des utilisateurs, afin de fournir la bonne solution pour ceux-ci. Un point fondamental pour la bonne conduite du projet et permettre une vraie adhésion des utilisateurs est de s'assurer que l'usage attendu est réaliste (qu'il colle à une réalité opérationnelle), que les parties prenantes sont alignées sur les critères de succès du projet (par exemple, être alignés sur le pourcentage de documents traités).

"Les nombreuses demandes exprimées par les équipes métiers sont l'illustration de la réussite de la démarche."

Une mise en application réussie à La Poste

L'exemple choisi pour illustrer le fonctionnement de bout en bout d'un projet *big data* et IA de traitement de documents en *back office* est volontairement simple et indépendant du contexte de l'entreprise, mais il a le mérite de décrire le fonctionnement technique du projet. À cela, il faut ajouter dans le contexte notamment d'une grande entreprise les complexités supplémentaires liées au silotage des *business units* (domaines d'activités stratégiques), des différents départements SI (systèmes d'information) et data parties prenantes dans l'entreprise, et les maturités hétérogènes des différentes équipes métiers.

La Poste a choisi d'investir massivement à la fois dans ses infrastructures et dans l'accompagnement complet de ces projets data & IA. Le groupe s'est ainsi doté d'un *data lake* (lac de données) unique pour la maison mère, complété par deux infrastructures *big data* – une pour la Banque et une pour la CNP. Parallèlement, afin de favoriser l'appropriation des sujets data par les utilisateurs finaux, La Poste a mis en place des équipes opérationnelles (IT et Data) ainsi que des équipes de conduite du changement et de valorisation des données. C'est cette structure solide qui a permis à l'entreprise de porter des usages novateurs et à fort retour sur investissement dans le traitement des documents en *back office* ; un premier usage en production permet déjà de traiter les factures entrantes de manière automatisée et de soulager les équipes comptables de tâches de saisie. La demande pour des usages du *data lake* a considérablement augmenté : nous comptons 200 usages en cours de réalisation pour les années 2022 et 2023. Indépendamment des ROI de ces usages, les nombreuses demandes exprimées par les équipes métiers sont pour moi l'illustration de la réussite de la démarche et la preuve que celle-ci répond à une vraie attente de nos équipes.