Projets data : comment réduire l'impact environnemental

En physique, l'entropie correspond à l'état de désorganisation de la matière. La minimisation de l'entropie permet l'organisation des éléments la plus éloignée de l'aléatoire. Elle est devenue un enjeu majeur de notre temps, qui concerne directement l'organisation des activités humaines. Et si nous l'appliquions à nos projets data ? Voici quelques pistes concrètes.

Sortir du paradigme « plus de data = plus de valeur »

La tendance humaine à vouloir produire et surtout à stocker plus de denrées trouve son origine dans nos comportements ancestraux pour anticiper les pénuries. Il semblerait que cette tendance se vérifie particulièrement dans notre gestion des données ! Par précaution, notre habitude est de prélever et de stocker le maximum de données pour maximiser la probabilité de n'avoir « rien oublié » lorsqu'il s'agira de démarrer l'analyse et l'utilisation de celles-ci. Cependant, avant toute accumulation de données, la question pourrait être désormais : avons-nous besoin de toutes ces données et pour quels usages projetés ?

Autrement dit, si une donnée n'a pas d'utilité aujourd'hui, sa valeur est en réalité négative car elle génère d'ores et déjà une dépense énergétique du fait de sa récolte, de son stockage et de sa circulation.

La donnée est une connaissance

La donnée se transmet, se diffuse, se copie vers des destinations multiples sans pour autant disparaître de son origine. Ainsi, le fait de générer une donnée n'est qu'une preuve de son coût et du potentiel multiplicatif de celui-ci et non de sa valeur.

Toutefois, toute expérience repose sur des hypothèses. Pour vérifier celles-ci, la génération de données est donc incontournable. Inutile cependant d'en générer ou d'en traiter plus que nécessaire !

Vérifier et évaluer l'intérêt d'une donnée

Les questions liées à l'intérêt tangible d'une donnée sont nombreuses :

– Quel objectif justifie la récolte et l'informatisation de cette donnée ?
– A quel ensemble de phénomènes une donnée permet-elle de contribuer ?
– Quelle donnée est la plus pertinente pour représenter au mieux un phénomène ?
– Quelle donnée est rendue obsolète par la récolte d'une nouvelle donnée ?

Il convient donc de s'assurer de la pertinence d'une donnée et de sa capacité à représenter un phénomène de manière fidèle à celui-ci mais aussi de déterminer les échelles d'intérêts

appropriées : à quelle fréquence, quel format, quelles transmissions, quelles duplications ou multiplications et enfin quelle infrastructure pour maximiser l'intérêt de cette donnée tout en minimisant son impact environnemental ?

Un volume de données « utiles » doit surtout être riche d'une diversité d'observations

Ce n'est pas parce que je dispose de beaucoup de données dans un champs précis que mes prévisions basées sur ces données seront meilleures. Pour atteindre des modèles plus performants, je vais avoir besoin de répétitions de situations (pour rendre mes conclusions significatives) mais également d'une diversité d'observations. Le trop plein d'informations sur une situation fixe induit notamment un phénomène de sur-apprentissage. Dans ce cas, la qualité de la prévision globale peut s'avérer diminuée par l'apport de données supplémentaires.

Répartir son « budget énergétique » selon la priorité des projets

Les coûts énergétiques liés au développement et à la mise en production d'algorithmes de machine learning peuvent s'avérer particulièrement consommateurs d'énergie. De la régression linéaire simple aux réseaux de neurones convolutifs, les besoins en ressources peuvent être facilement décuplés, ce qui n'est pas systématiquement souhaitable. Aussi, le choix de l'algorithme devra être considéré en fonction du gain de précision qu'il apporte, au regard des ressources nécessaires et en fonction du niveau de priorité du projet.

La modélisation s'avère parfois être la compensation d'un manque de connectivité entre systèmes d'information

Si tant de données sont enregistrées et traitées par les entreprises, ce n'est pas souvent dans un but de modélisation statistique mais bel et bien dans l'objectif d'un service personnalisé et directement connecté. La modélisation trouve son intérêt dans la généralisation d'un phénomène et ne doit pas être confondue avec la connexion directe de systèmes d'information.

Prenons un exemple simple. Le restaurant collectif d'un immeuble de bureau souhaite prévoir au mieux le nombre de convives à déjeuner le jour j et les assortiments de repas à prévoir. Il déballera et préparera le matin même le nombre et la variété des repas prévus en incluant une marge d'erreur dans l'objectif de pouvoir satisfaire l'ensemble des convives, y compris ceux dont les comportements sont les plus difficiles à anticiper. Pour y parvenir, le restaurant a deux options.

Option 1 : il peut développer un ou plusieurs modèles de prévision basés sur l'observation des repas historiques et des données corrélées à ces observations telles que la météo. Il peut également s'entendre avec les entreprises de l'immeuble pour récolter un certain nombre d'informations impactant le nombre et la typologie des convives (présence des salariés sur le lieu de travail, formations externes…) et ainsi améliorer la qualité de ses prévisions.

Option 2 : il peut développer une application permettant aux salariés des entreprises présentes dans l'immeuble de communiquer à la fois leur venue et leurs choix, jusqu'au jour même.

Ainsi, la première option repose sur des modèles statistiques probabilistes dont la précision varie

en fonction de la capacité à récupérer les données corrélées avec l'activité. C'est-à-dire la capacité à connecter différents systèmes d'informations entre eux (communication anonymisée des agendas électroniques des employés des différentes entreprises de l'immeuble).

La seconde option vient alimenter directement le système d'information du restaurant et garantit ainsi une optimisation de ses besoins tout en minimisant le gaspillage. La modélisation statistique peut alors trouver sa place dans la prévision de la consommation à plus long terme dans un objectif d'optimisation de ses stocks.

Dans la réalité, il est encore rare de constater de telles connexions directes d'évaluation de l'offre directement par la demande. Les modèles statistiques viennent en quelque sorte pallier ce manque d'outils déclaratifs ou de connexions entre systèmes d'informations.

La structuration de la data doit être drivée par les usages

Lorsqu'une donnée est utilisée à objectif d'analyse ou de modélisation, elle est au préalable étudiée, nettoyée et surtout préparée afin de pouvoir communiquer avec d'autres données. L'exemple le plus courant de préparation est la mise à la même échelle de temps d'un ensemble de données. Ce n'est qu'alors que ces données peuvent « communiquer » et que nous pouvons déduire une relation entre ces variables ou ces phénomènes.

Or, sur de multiples projets, les mêmes sources de données sont préparées en fonction des autres sources et de leur granularité respective. Ainsi, il n'est pas rare de trouver dans un même lac de données, la répétition des étapes mentionnées ci-dessus sur une même source de données. Dans les faits, davantage de travaux peuvent être mutualisés et nous pourrions éviter la redondance de certaines transformations et stockages.

Les clés pour éviter cette redondance reposent sur :

- Un partage des pipelines : cycles visuels de chargements, de préparation et de transformations de données liées à un projet permettant la réutilisation de tout ou partie de ceux-ci.
- Un accès intelligent aux éléments de documentation (intelligent search) et de visualisation des projets.
- Une analyse des interactions entre projets menée en continue.

S'ils sont bien utilisés, les nouveaux outils de design de pipelines et de construction et de suivi de projets de type DataOps et MLOps permettent de répondre au moins en partie à ces objectifs.

Échantillonnage et sondage de données : quel volume pour quoi faire ?

Pour chaque analyse lancée ou chaque exercice de modélisation, il convient de se demander quel volume de données est nécessaire pour établir une approximation suffisante de ce que l'on

souhaite démontrer. Pour prendre une décision, nous n'avons que très rarement besoin d'une analyse précise à la décimale près. Selon les scénarios, parfois 10 % des données suffisent à établir le constat souhaité et à suivre les évolutions d'un phénomène.

Les techniques de sondages appliquées aux populations peuvent être appliquées aux données : tirages aléatoires simples, stratifiés ou en grappes. Selon les cas, une méthode peut être utilisée afin d'estimer un résultat en fonction du niveau de précision souhaité.

A la différence des sondages politiques, les techniques de sondages appliquées aux données permettent non pas d'économiser le coût lié à la récolte de ces données mais d'économiser une partie des ressources liées à leur traitement. Tout comme dans le cas des sondages classiques, le niveau de précision souhaité détermine le taux de sondage nécessaire pour y parvenir et un intervalle de confiance (ou marge d'erreur) est indiqué.

Prenons l'exemple de la latence des tableaux de bords. Étant donné le volume de données désormais à disposition pour quantité de projets, il n'est pas rare de voir des tableaux de bord qui « rament » au chargement. Parfois, l'erreur est de vouloir y connecter un volume trop important de données (agrégées ou non) et l'on cherche donc la solution côté infrastructure, mobilisant de nouvelles ressources informatiques, alors qu'une sous-partie de ces données sélectionnées intelligemment suffirait à régler ce problème de latence, à iso-ressources.

L'utilité marginale de la donnée

La quantité de précisions apportées par une unité supplémentaire d'information correspond à l'utilité marginale de la donnée. Le gain peut être évalué en regard de l'énergie incrémentale nécessaire pour son traitement. Pour chaque scénario, il existe un seuil à partir duquel l'apport en précision d'une donnée supplémentaire est nul, proche de 0, ou même négatif. Il n'y a alors plus d'intérêt à récolter ou utiliser cette donnée.

Il existe plusieurs scénarios d'application de cet indicateur afin d'arbitrer sur l'utilisation de données supplémentaires :

- Au niveau de la granularité de production et de transmission d'une information : supposons par exemple un système IoT qui produit et transmet une information de température toutes les secondes alors que l'utilisation faite de cette température n'est établie qu'au niveau de la minute.
- Au niveau de l'obsolescence : est-ce que l'utilisation d'une donnée historique supplémentaire plus éloignée améliore ou dégrade la qualité de ma prévision ? On parle alors de prédictibilité marginale négative de la donnée.
- Au niveau du volume de données estimé nécessaire en fonction du niveau de précision souhaité (voir paragraphe précédent sur l'échantillonnage et le sondage).

Ainsi, la production, la transmission et l'utilisation d'une donnée doit toujours être calibrée en regard d'un objectif, d'un besoin et du niveau de précision souhaité.

Conclusion : la nécessité des approches frugales

La montée en puissance du Big Data crée parfois des environnements dont la complexité et la multiplicité limitent la maîtrise de la structuration et de l'utilisation des données. Les ressources matérielles et énergétiques sous-jacentes sont souvent très importantes.

Etant donné les volumes de données à disposition, il est primordial de développer des techniques de minimisation des coûts énergétiques liés à l'implémentation de ces projets. D'autant que les résultats de ces recherches aboutissent à la mise en place de nouveaux services qui se diffusent très rapidement et en masse.

La modélisation statistique dont nous parlons beaucoup de manière indirecte aujourd'hui lorsque nous évoquons l'intelligence artificielle ou la Data Science est avant tout un ensemble de méthodes d'approximations. Prenons l'exemple des prévisions météo. On ne peut qu'estimer une situation météorologique future et non pas la prévoir avec exactitude.

Le volume de données pour parvenir à une estimation est certes important mais au-delà d'un certain seuil, le gain apporté par une donnée supplémentaire devient décroissant. Il faut donc des approches frugales pour ne pas générer plus de données que nécessaire.

Au-delà de la documentation classique, les nouveaux outils permettant de bâtir et de partager les pipelines de projets devraient permettre un meilleur partage des transformations de données déjà établies dans les systèmes d'information et d'éviter les redondances liées par exemple à la préparation de ces données.