Gawaboumga

Geste de savoir

De Qlik à Power BI : rétrospective métier

mardi 27 août 2024

Table des matières

Introduction

Il y a 6 mois, j'écrivais un billet sur la mise en place d'un data lake afin de répondre aux besoins de reporting et de data gouvernance. Il est temps de revenir une première fois sur la situation, sur ce qui a fonctionné, ce qui n'a pas réussi et les points qu'il reste à améliorer. Dans ce billet, on s'intéressera principalement aux aspects métiers (business) de la problématique. Vous pouvez consulter les aspects techniques dans cet autre billet .

1. Données de référence et données de référence ?

Une difficulté rencontrée lors de l'exercice a été la réalisation qu'il existait plusieurs types de données de référence, résultant des différents objectifs que l'on souhaite réaliser :

— Les données analytiques : Ce sont les données qui servent à des usages analytiques, que ce soit par une exploration directe par les *data analysts* (au travers de SQL, Python, Spark, …) ou plus indirectement par les outils de *Business Intelligence*. Les gens du métier n'interviennent que très rarement de manière directe avec les données, ils passent systématiquement par des solutions informatiques. Néanmoins, des *spreadsheets* Excel sont assez classiques et il est parfois nécessaire de contraindre les types des colonnes ainsi que l'ensemble des valeurs possibles pour un ensemble de cellules afin d'éviter des soucis de traitement.

— Les données de référence : Ce sont des données qui peuvent servir à des usages analytiques, mais qui ont souvent une complexité intrinsèque qui les rendent difficiles à exploiter directement. Elles sont davantage adressées en tant que service et non comme source directe.
Prenons les activités économiques d'une entreprise, celles-ci peuvent être multiples, avec une certaine classification (activité primaire, secondaire & tertiaire), des périodes de validité, … alors que, dans le fond, on veut peut-être juste savoir s'il s'agit d'une banque ou d'un petit commerçant de quartier.
Ces données peuvent également avoir un volume particulièrement grand, quel sera l'impact

sur les performances dans des outils d'analyse ?
Finalement, il peut être intéressant de proposer ces données comme un véritable "service" au sens informatique (API - webservice), en proposant des approches ponctuelles ou par paquets (*batch*), et ce, par une variété de moyens de communication afin de minimiser les frictions éventuelles.
L'accès aux données se fait le plus souvent de manière automatisée par des systèmes informatiques, mais les personnes du métier peuvent éventuellement interagir directement avec elles, afin de corriger une donnée ou simplement la consulter.

— Les données de référence ponctuelles : Certaines données de référence sont difficilement automatisables (ex : des emails récapitulatifs provenant d'autres institutions ou entreprises) que ce soit par leur variété intrinsèque de format (demande souvent informelle) ou une complexité monstrueuse pour finalement un intérêt plus que limité. Il est alors préférable d'offrir la possibilité aux gens du métier d'encoder directement ce genre d'informations. Il peut également y avoir un aspect de résultats issus d'un long processus d'analyse, difficilement synthétisable.

— Les dimensions de référence : Ce sont des données qui n'ont généralement qu'une utilité purement métier. Typiquement, faire correspondre des identifiants techniques (*enumeration*) à des notions métiers ou manipuler ces notions sous un autre aspect, plus adapté au besoin présent (regrouper des catégories). Ici, la relation est dans l'autre sens, ce sont les gens du métier qui manipulent directement cette information (qui la modifie, complète ou supprime), et qui est alors exploitée par les systèmes informatiques.

Ces différentes notions se sont traduites par des processus métier distincts. Des questions complexes émergent : qui est responsable des modifications, comment proposer une gestion unifiée des droits d'accès au sein de différents outils, quid de l'historique des modifications (données historiques - comme à l'époque vs historisées - comme à l'époque mais avec les informations actuelles), de l'interaction avec les systèmes informatisés (il faut parfois prévenir le côté technique des changements qui vont être apportés), de la communication des données ponctuelles qui sont transversales.

2. Formation & accompagnement

L'organisation de formation à Power BI se réalise difficilement dans un contexte où le nombre de participants est élevé. En effet, si la manipulation sommaire de l'outil et de ses concepts est relativement intuitif et rencontre peu de difficultés. Il est facile de tomber sur un grand nombre de questions, quelles soient techniques (comment on fait X ?) ou analytiques (et qu'est-ce qui se passe si on compare X et Y ?) mais globalement, les choses se passent plutôt bien. Là où davantage de difficultés sont rencontrées, c'est lors des formations plus pointues :

1. Il est préférable de travailler avec un seul groupe de travail, d'autres membres pourraient présenter des signes de désintéressement parce que cela ne concerne pas "leurs" données alors qu'ils seront confrontés aux mêmes difficultés …
2. Il est important d'incorporer des nombreux exercices pratiques entre-coupés des principes généraux qui gouvernent l'outil afin qu'ils se confrontent aux problématiques classiques une première fois et apprennent comment les résoudre.

Il est bien présomptueux de croire qu'une formation suffira à la montée en compétences des différents participants. C'est pour ça qu'il est intéressant de proposer un service d'accompagnement qui se traduit par des phénomènes plus ou moins officieux tels que s'asseoir à côté de la personne, des réunions sur *teams*, la formation de groupes de travail / discussion, des permanences, … Si je devais nommer des problématiques qui me viennent en tête : On se sent parfois limité pour faire de l'analyse exploratoire des données :

— Des comportements peu intuitifs au niveau des relations : bi-directionnelle, cycles, actives/inactives ;
— Que ce soit par ce langage de programmation Power Query M qui, malgré une certaine simplicité, est difficile à appréhender par les gens issus du métier ;
— Difficulté pour faire des requêtes un peu complexes (*NOT EXISTS*) ;
— Problèmes de performances liés au volume de données ;
— Les noms "métier" qui ne correspondent pas forcément entre l'applicatif, le stockage et le métier, ce qui introduit de la confusion. Problématique péniblement résolue par le catalogue de la donnée et c'est assez fastidieux de documenter au sein de Power BI ;

Ou dans l'usage au quotidien :

— Difficile d'entretenir plusieurs types de compte pour se connecter à une même source de données ;
— Impossibilité de se connecter à plusieurs 'Power BI datasets' ;
— Capacité de maintenir plusieurs versions traduites (internationalisation, localisation) plus que sommaire ;
— Gestion des modèles et des rapports plus complexes qu'il n'y paraît en terme des modifications apportées ;
— Difficulté pour intégrer des données programmatiquement (nous avons un catalogue où le métier peut compléter les définitions des différentes données, il serait agréable de pouvoir modifier ces informations au sein de Power BI sans faire des acrobaties …) ;
— Comment fait-on pour regrouper les mesures dans un même dossier ? Ha oui, j'avais oublié ;
— Mise-à-jour sur une heure UTC et non locale. Merci les changements d'heure ;

Malgré tout, Power BI reste relativement intuitif et puissant. Mais, sans accompagnements ou formations, il est facile de se retrouver coincé et frustré. Or, l'objectif est d'éviter au maximum les appréhensions face à l'outil afin que les gens du métier gagnent en contrôle sur leurs données et les questions qu'ils peuvent poser dessus. *Easy to learn, hard to master.*

3. Communication auprès des utilisateurs

Sans doute l'un des gros points d'amélioration à opérer. Tant avec l'augmentation du nombre d'utilisateurs en interne qu'avec la complexification des besoins et processus, il est important de communiquer des changements qui sont opérés. Plusieurs solutions sont envisagées afin de répondre à ce problème :

— Des notes de version (*release notes*) incluant des changements au sein des données (par exemple, nous avons décidé de normaliser les numéros d'entreprise Belge au format de la Banque Carrefour des Entreprises : 0252.796.351) ;
— Les nouveaux jeux de données intégrés ;

- La modification ou la création de processus métier (ex : nous avons rajouté une vue unifiée du catalogue de données permettant une recherche d'un terme) ;
- La collecte de besoins et raffinement des concepts (ex : nous avons travaillé sur la notion de lineage, comment présenter cette information de manière digeste ?) ;
- La création d'un site web reprenant les différentes informations ainsi que d'éventuelles procédures (ex : mettre à jour un fichier par un utilisateur) ;

Là où le bas blesse davantage, c'est sur le fait que nous n'avons pas encore mis en place un contrôle qualité et une assurance qualité sur les données. S'il y a le moindre problème, cela passera sous le tapis et personne ne sera mis au courant. Heureusement, nous arrivons à un stade de maturer où ces problématiques peuvent être envisagées de manière plus extensive que ce que nous proposons actuellement (validation plus que sommaire).

Étonnamment, il y a plutôt une bonne adhérence aux pratiques de gouvernance. Certaines personnes sont même plutôt pro-actives dans leurs démarches et les défendent auprès des leurs !

4. Site web - Data catalog

Le site web reste une bonne idée, les utilisateurs parviennent à leurs fins sans trop de subtilités d'utilisation. Le tout a été réalisé malgré un manque de compétences évidentes en développement web en interne. Parmi les nouvelles fonctionnalités qui présentent un intérêt certain, on retrouve :

- Authentification au travers de Windows SSO (SAML) ;
- Recherche des champs dans le data catalog ;
- Visualisation de l'affiliation des données (data lineage) ;
- Extension de la configuration pour le fonctionnement du data lake ;
- Visualisation des relations qu'entretiennent les différents data sets entre eux ;
- Définitions des dimensions de références ;
- Encodage des données de référence ponctuelles ;
- Enregistrement d'utilisateurs techniques sur des données de références ;
- Possibilité de relancer certains jeux de données (par exemple, suite à la mise à jour de fichiers utilisateurs) ;
- Visualisation de la dernière mise-à-jour des données ;

La problématique principale réside surtout dans le fait que le site actuel soit moche (et pas encore entièrement traduit).

Conclusion

En conclusion, on arrive à un stade de maturité déjà nettement plus avancé. Ici, la préoccupation première s'orientera vers davantage de polissage ; que ce soit des améliorations graphiques pour les différents outils ou la correction de bugs. Il reste évidemment de grandes thématiques à entreprendre :

- Contrôle qualité (QC) et assurance qualité (QA) ;
- Remédiation de la donnée (data remediation) ;

- Résolution des entités (entity resolution & master data management) sur des aspects davantage métiers. Nous avons déjà bien avancé sur le côté technique de la question ;
- Politique & stratégie de documentation des jeux de données accessibles par le public - Open Data (ex : information sur les dépendances fonctionnelles des champs) ;
- La mise en place d'une stratégie liée à l'intelligence artificielle ;
- Les interactions avec d'autres systèmes (notamment avec la suite "Power Platform") ;
- Documentation des flux existants et inventaire de la donnée ;

Bref, beaucoup de travail prévu pour les prochaines années. Ce qui est chouette, c'est qu'on quitte la longue étape de mise-en-place du produit et qu'on commence enfin à s'attarder sur des thématiques plus matures. C'est aussi l'occasion de prendre une place plus importante dans les discussions autour de nouveaux projets sur la manière dont les données pourront s'intégrer aux flux existants.