Gawaboumga

Geste de savoir

De Qlik à Power BI

mercredi 20 mars 2024

Table des matières

Introduction

Ces derniers mois, j'ai travaillé sur la mise en place d'un data lake afin de répondre aux besoins de reporting et de data gouvernance. L'objectif global est plus global qu'une simple migration de Qlik (et Business object) vers Power BI. Je souhaiterais partager les problématiques et solutions que j'ai rencontrées et proposées afin d'améliorer la collaboration au sein de mon job actuel.

Qlik Sense et Power BI sont deux outils populaires dans le domaine de la *business intelligence* (BI). Ces outils servent principalement à aider à l'identification, au développement et à la création de nouveaux objectifs stratégiques. Ils se veulent généralement assez accessibles dans leur approche, en permettant d'analyser des données sous de nombreux angles.

Mais dans le fond, pourquoi migrer d'un outil à l'autre ? De nombreux éléments peuvent rentrer en jeu dans ce genre de considérations :

— Coûts et allocations des ressources : Quels sont les coûts initiaux et continus de ces technologies ? Est-ce que l'investissement en ressources humaines compense-t-il cette dépense ? Quid du support à long-terme, tant de la part de l'IT, que des gens associés au domaine (data analyst) ?
— Formation, accompagnement & communication des utilisateurs : Comment organiser des formations afin d'aider l'adoption au sein des utilisateurs ? Que peut-on mettre en place afin d'aider à cette transition et ainsi encourager les équipes ? Comment rassembler les retours d'expériences et garder les différentes partie-prenantes impliquées ?
— D'un point de vue technique : Est-ce que cela va diminuer des problèmes de *downtime* (coupure d'intermittence), de *data loss* (pertes de données) ou de *disruptions* (perturbations) ? Comment les performances des deux outils sont comparables ? Est-ce que cela permettra de mieux gérer la croissance de l'entreprise ? Comment convertir les flux actuels de données pour ce nouvel outil ? Existe-t-il des intégrations spécialisées ou des outils de conversion ?
— Alignement des objectifs & data governance : Faut-il profiter de cette migration pour réaligner les objectifs ? Peut-on effectuer un nettoyage de printemps ? Ne serait-ce pas l'occasion de compléter la data governance en revoyant les différents processus ?

L'élément le plus crucial dans cette migration réside dans la définition d'une stratégie à long-terme pour les besoins de business intelligence. C'est un changement profond dans la société puisqu'elle touche aux flux de communication. Serait-ce l'occasion de s'attaquer à des problèmes plus larges comme de l'architecture d'entreprise ?

1. Situation actuelle et objectifs

Les facteurs ayant poussé à la migration de Qlik Sense à Power BI ont été multiples. Ils s'inscrivent avant tout une politique de modernisation, ce qui permet de faire des économies puisqu'il n'y aura plus qu'une seule technologie à maintenir.

La mutualisation se fait également en termes de connaissances et de ressources humaines. Il est plus simple de trouver des gens spécialisés dans cette technologie que dans les nombreuses autres que l'on peut retrouver au niveau des différentes divisions. Power BI s'intègre également mieux avec les autres produits Microsoft et notamment autour de l'environnement 'Azure' (autour de Fabric). Microsoft a clairement décidé d'investir davantage dans Power BI, en offrant des mises-à-jour très régulières de leur produit.

Ce changement s'opère également dans une remise en question des processus afin de repartir sur des bases plus saines. Cela s'intègre parfaitement dans une politique d'architecture d'entreprise de meilleures collaborations des différentes unités, tout en veillant à mieux respecter les différentes législations et régulations. D'aucuns peuvent profiter de ces changements de technologies à fin de revoir les processus de fond en comble et de mises-à-jour des processus business. En autres, afin de réduire la part de *shadow it*. Une certaine remise en question est souvent nécessaire, tant par la longévité (est-ce que le business d'il y a 20 ans est le même que les besoins actuels ?), que la fonction 'non-technologique' (le cœur de métier est n'est pas toujours la technologie, mais celle-ci a gagné en importance) ou la croissance soudaine (les problèmes que l'on rencontre évoluent en fonction de la taille de la compagnie).

De nouveaux cadres légaux et réglementations font également bouger les choses et poussent à une collaboration plus étroite des différentes entités. Il est souvent plus simple de créer quelque chose de neuf que de migrer l'existant. Certains poussent pour définir des notions de *maturity model*, mais vu les différents éléments vus, la conclusion est souvent assez évidente …

2. Data gouvernance

Afin de démarrer ce type de travaux, il est intéressant de faire un post-mortem, comprendre quelles difficultés sont rencontrées actuellement et que souhaite-t-on obtenir au final. On essaye donc de formuler une *data strategy* facilement compréhensible et alignée avec les objectifs organisationnels. Cela se traduit par l'identification des jeux de données 'critiques' au business, à l'analyse de méthodes de travail liées à la gouvernance de la donnée et établir des balbutiements d'architecture. La gouvernance des données et la conformité (*compliance*) vont souvent de pair. Développer des processus et procédures afin d'assurer l'intégrité, la sécurité, la durée de vie et la conformité des données en accord avec les régulations est un enjeu capital.

Mais cela ne peut s'effectuer sans une très forte collaboration au sein de l'entreprise, que ce soit avec l'équipe légale, ou des parties-prenantes expertes de leur domaine et donc souvent au

courant des règles qui les concernent, et ainsi aider à prioritiser les opérations. Cette collaboration est l'occasion de mettre en place des règles pratiques et d'encadrement afin de promouvoir l'innovation et la culture de la donnée. Partager les retours d'expérience au sein de discussions planifiées est souvent pénible, peut-être vaut-il mieux convoquer une réunion afin de présenter ce qu'on a en tête et de recevoir les critiques ? Il peut également être intéressant de travailler directement avec ceux à l'origine du *shadow it*, soit en fournissant des personnes dédiées à des petits projets informatiques, soit en les intégrant de manière plus importantes (et prioritaires) dans ce nouveau processus (*Center of Excellence*). Cela permet de semer petit à petit une culture de la donnée au travers de l'organisation en montrant ce qu'il est possible de réaliser et de montrer comment cela améliore leur quotidien.

Outre ces aspects purement humains, que peut-on mettre en place d'un point de vue informatique afin d'assister à cette transition ?

Bien sûr, au-delà d'une participation étroite aux notions précédemment mentionnées, ils sont avant tout là pour proposer la moindre résistance possible au travail des analystes. Mise en place d'un catalogue de la donnée et d'un système de métadonnées. Cela permet d'obtenir une meilleure transparence tant pour identifier des risques (Que se passe-t-il si ce serveur meurt ? Qui a accès à cette donnée ?), que des aspects plus pratiques comme l'affiliation (*data lineage*) ou les définitions. Établir des mécanismes de contrôle des accès, de sécurité (*encryption*), de qualité de la donnée (QA - qualitatif & QC - quantitatif) ou d'audit que ces derniers sont bel et bien effectués. Simplifier l'intégration des données de référence dans les différents systèmes (*Data Governance Workflow Automation*) et offrir des performances satisfaisantes pour tous les types de clients. Il faut veiller à faire une veille des tendances de l'industrie et des technologies qui émergent afin d'éventuellement pouvoir faire ressortir de nouvelles possibilités sur les données.

3. Un site web

Afin de répondre à une partie de ces problématiques, développer un site web afin de servir d'interface entre les différents utilisateurs métiers et la partie plus technique. Celui-ci s'oriente autour de différents concepts :

- Data source : les sources de données originelles où l'on va puiser les informations, qu'elles soient internes (serveur) ou externes (ftp/site web).
- Data set : les jeux de données représentent l'information consolidée depuis plusieurs data sources ; ce sont eux qui sont accessibles aux différents utilisateurs (et Power BI).
- Responsability matrix : afin de définir une matrice RACI des responsabilités, des utilisateurs sont définis et sont ensuite associés aux différents concepts avec des rôles spécifiques (*data owner*, *data steward*, …).
- Glossary & Metadata : les utilisateurs peuvent ajouter différentes métadonnées ainsi que compléter un glossaire permettant tant de décrire brièvement le jeu de données, les différentes tables ou même les champs qui le composent (avec des métadonnées éventuelles).
- User interactions : les utilisateurs peuvent également interagir avec le système :
  - Par l'import de fichiers nécessaires à leur processus (toutes les spreadsheets excel ne peuvent pas être automatisées), elles sont automatiquement validées dès leur ingestion par rapport à ce qui a été défini entre l'équipe technique et business ;

— Par la définition de transformations ad-hoc (*custom*) en intégrant des jupyter notebook respectant un certain cadre ;
— Par la production d'exports personnalisés où un email de notification sera envoyé au(x) destinataire(s).

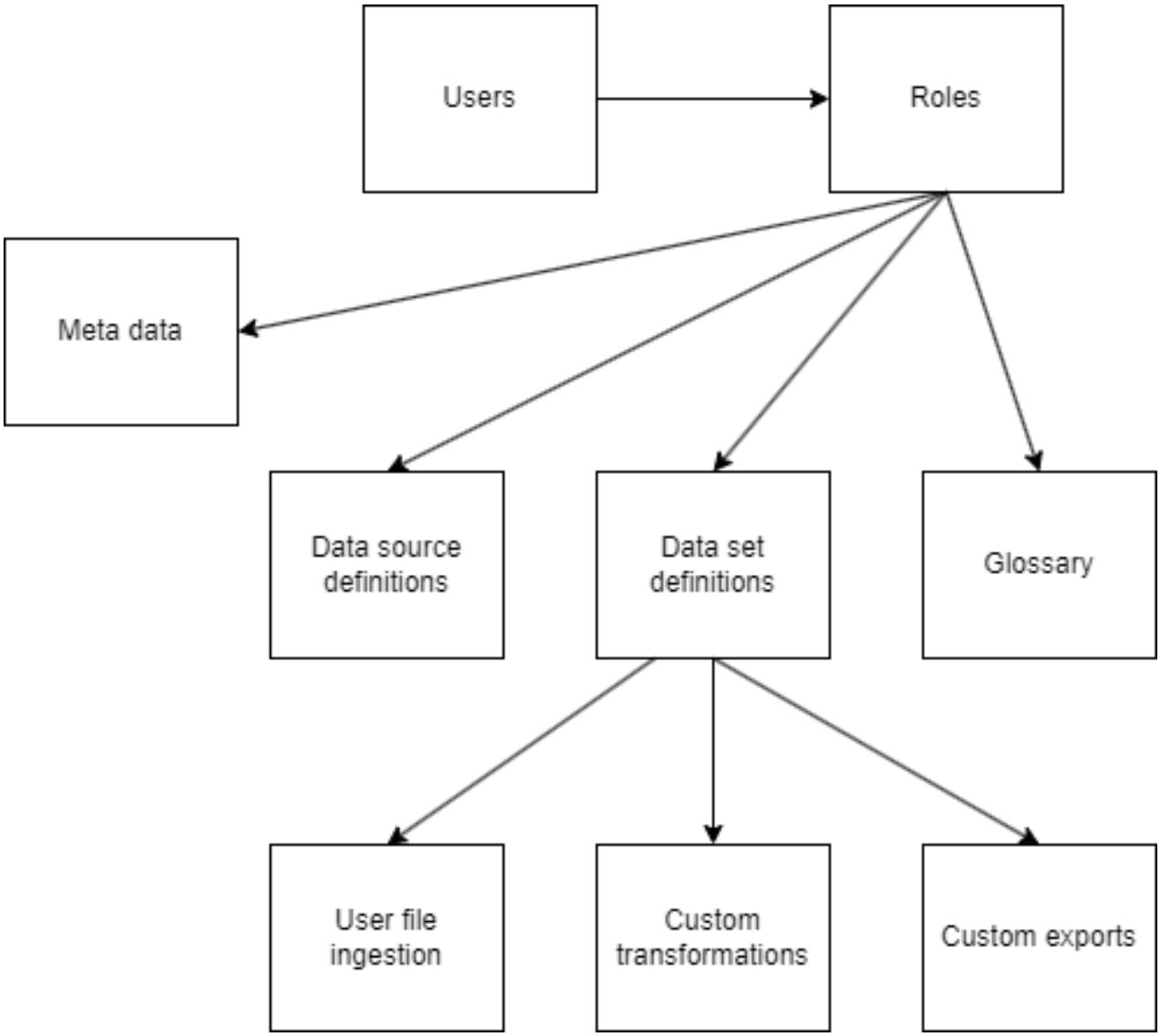

FIGURE 3.1. – Le web serveur permet aux gens du métier d'interagir directement avec le data warehouse tout en remplissant les besoins de data gouvernance

4. Technologies

Pour les technologies qui composent la partie technique, on retrouve des grands classiques des processus d'ETL : *Apache Airflow*, *DBT*, *Snowflake*, *Azure*, *Fabric*, *pandas* en Python … Ces choix sont plutôt classiques (*nobody ever got fired for buying IBM*) mais font clairement le travail ; les clusters spark ont tendance à être fort coûteux et sont assez difficile à maîtriser. On peut néanmoins émettre des avis plus réservés sur certaines technologies :

— Python n'est pas connu pour sa vitesse, mais rien n'empêche d'appeler une transformation dans un autre langage depuis Python. Il est également facilement accessible et permet éventuellement aux data analystes d'écrire leur propre transformation !
— Airflow, sans plus, je regrette le fait qu'on ne puisse définir différents ordonnancements. Je souhaiterais qu'un processus tourne quotidiennement entre Novembre et Février et puis une fois par semaine les mois restants.
— DBT …

4.0.1. DBT

J'ai des grands problèmes avec cette technologie parce qu'elle se positionne de manière très étrange pour le moment. Le principe consiste à écrire les transformations sous la forme de SQL. Ces transformations sont ensuite composées sous la forme d'un arbre de dépendance et exécutées. Le problème est qu'il ne s'agit pas d'une abstraction sur SQL, donc on retrouve toutes les difficultés auxquelles s'attendre : performances, traitement de texte inadapté, lié à une base de données en particulier, support pauvre des éditeurs de texte et impossible de renommer des variables sans modifier les 36 scripts qui en dépendent … Au final, je me suis retrouvé à écrire un DSL (*domain specific language*) qui me permet de générer tous les scripts (SQL et YAML) nécessaires au bon fonctionnement de DBT, et donc, pourquoi ne pas appliquer le SQL moi-même et supprimer DBT au final ? Ce qui est quand même positif, est que l'on ne se concentre plus que sur la transformation en tant que telle et les besoins des utilisateurs finaux et trouver des gens métiers ayant une connaissance sommaire de SQL n'est pas si rare.

Mais étonnamment, la création de ce DSL a eu l'avantage de permettre une intégration beaucoup plus fine des processus :

— Rajouter une source de données consiste juste à cliquer sur 3 boutons dans une interface graphique qui me produit une description des tables d'origine dans mon DSL.
— On peut facilement exporter l'affiliation des données (*lineage*) s'il n'y a pas de transformations spécifiques dessus (en Python).
— On peut mettre à jour le glossaire disponible aux utilisateurs automatiquement.
— Cela génère également automatiquement des contrats au niveau des données reprenant tant leur type, que leur nullabilité, intégrité référentielle, … qui seront ensuite employés lors de la mise-à-jour des sources de données afin de s'assurer qu'il n'y ait pas eu changements.

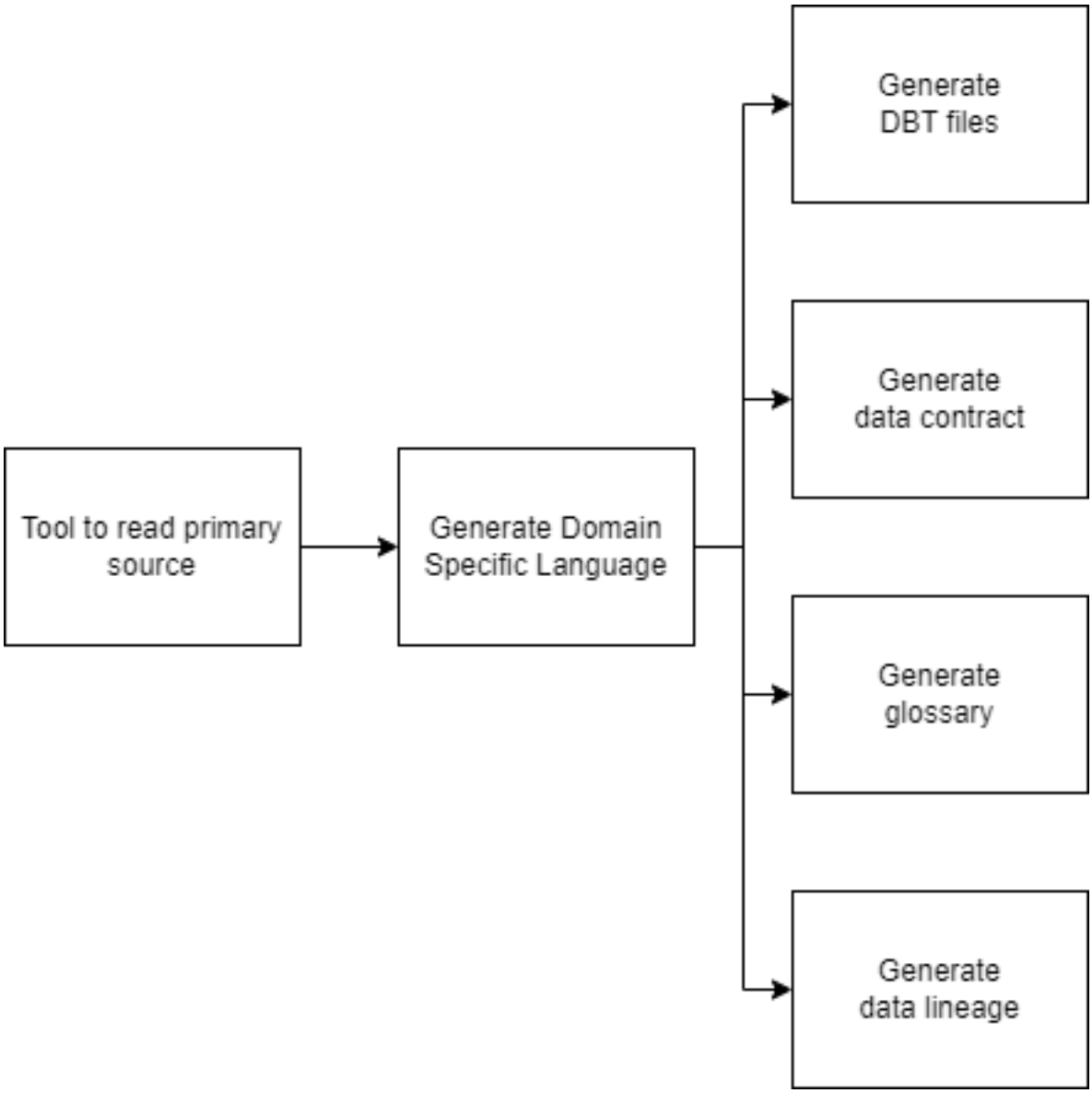

FIGURE 4.2. – Le DSL permet d'accroître les fonctionnalités de DBT afin de compléter les besoins et fournir une meilleure intégration de bout en bout

5. Data lake / data warehouse

Pour ordonnancer les tâches définies au sein du site web, on emploie Apache Airflow qui va chercher les définitions des jeux et sources de données afin de savoir ce qui doit être effectué. Une fois les sources de données mises-à-jour, on applique les différentes transformations possibles, qu'elles soient en Python ou DBT et fournies par des utilisateurs métiers ou techniques. On exporte ensuite éventuellement des données si cela est nécessaire et on met à jour les permissions d'accès par rapport aux rôles et responsabilités des différents utilisateurs.

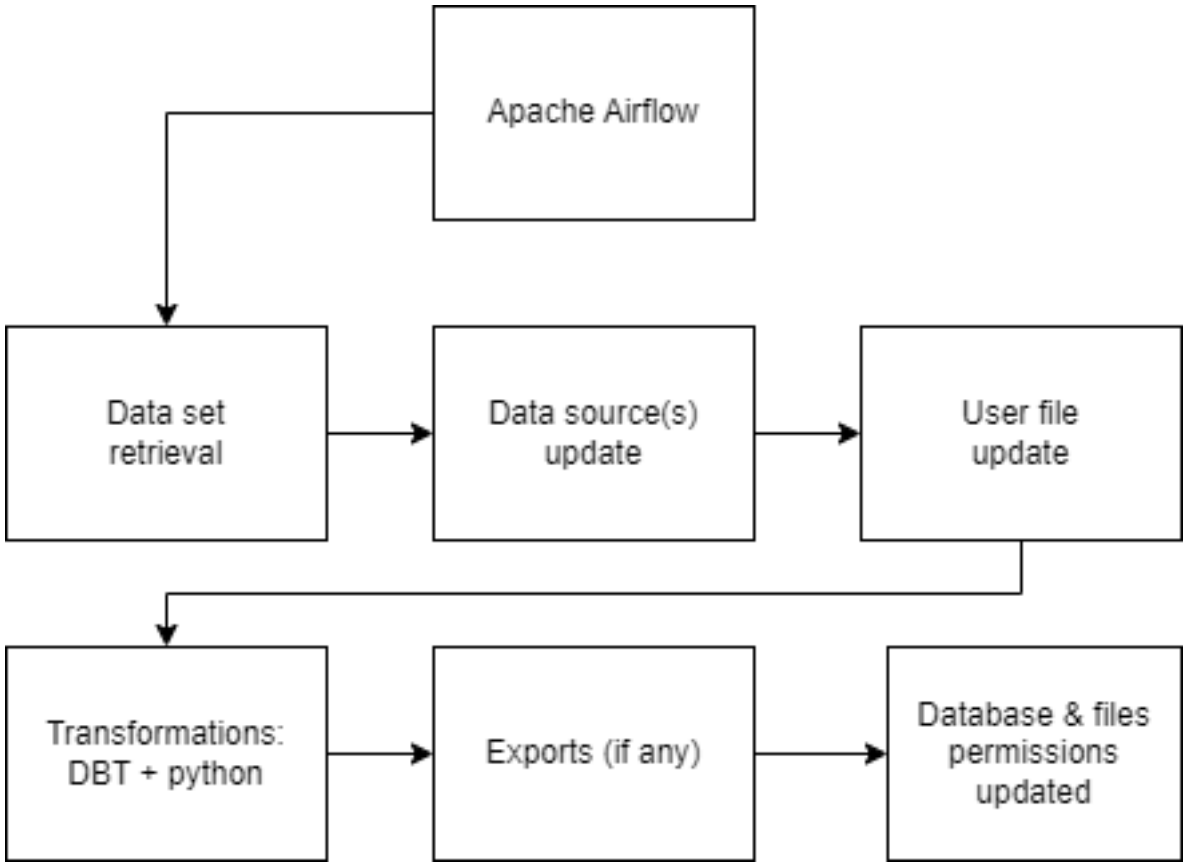

FIGURE 5.3. – Principe général de mise-à-jour d'un jeu de données au sein du data lake

Des considérations additionnelles ont été prises en compte dans la création de ce data lake/warehouse. En effet, au sein de l'entreprise, les notions d'écologie sont très importantes et se traduisent par l'inclusion et la sobriété (numérique) :

- Inclusion : parce que le site web a été développé en accord avec les réglementations WCAG 2.1. Transparence totale des processus et support continu aux différents utilisateurs métiers.
- Sobriété numérique : réduire son empreinte carbone permet de diminuer les coûts.
  - En effet, de par la création d'un data warehouse, il n'existera plus qu'une seule source authentique de données et non une multitude de copies au sein de l'entreprise.
  - Compression des données qui ne nécessitent que des usages occasionnels.
  - On stocke au maximum des fichiers afin d'éviter d'avoir une grosse base de données qui est allumée en continu.
  - Les serveurs de travail (*workers*) sont allumés uniquement sur demande, à l'apparition d'une nouvelle tâche.
  - On préfère employer des vues, et on ne matérialise les tables qu'en cas de nécessité.

Conclusion

En six mois, beaucoup de choses ont déjà pu être mises en place, mais le *backlog* a grandit également à une vitesse tout aussi impressionnante …

Quels sont les objectifs suivants ?

— Finir la conversion de tous les flux de données existants.
— Finir l'identification des jeux de données de référence.
— Améliorer les formations disponibles à ce sujet en proposant des ateliers pratiques où les data analystes apprennent à manier leurs données.
— Créer des groupes de discussions, où les gens peuvent demander des avis externes sur ce qu'ils ont en tête.
— Mise en place de processus de qualité (QA & QC) ainsi que de correction (*data remediation*).
— Aide aux besoins prédictifs et liés à l'intelligence artificielle.
— Travailler sur les problèmes d'architecture de la donnée.