CHAPITRE 2: INTEGRATION DES DONNEES

I. Définition

L'intégration des données est le processus qui consiste à combiner des données provenant de différentes sources dans une vue unifiée : de l'importation au nettoyage en passant par le mapping et la transformation dans un gissement cible, pour finalement rendre les données plus exploitables et plus utiles pour les utilisateurs qui les consultent. Les entreprises sont en train de mettre en place des initiatives d'intégration de leurs données pour les analyser et les exploiter plus efficacement, en particulier face à l'explosion des données entrantes et l'arrivée de nouvelles technologies comme le cloud et les big data. L'intégration de données est une nécessité pour les entreprises innovantes qui souhaitent améliorer leur prise de décision stratégique et augmenter leur avantage concurrentiel.

En matière d'intégration des données, il n'existe pas d'approche universelle ou standard. Toutefois, les solutions d'intégration de données partagent généralement quelques éléments, dont un réseau de sources de données, un serveur maître et des clients qui accèdent aux données à partir de ce serveur maître.

Dans la plupart des processus d'intégration des données, le client envoie une demande de données (requête) au serveur maître. Le serveur maître importe les données nécessaires à partir de sources internes et externes. Les données requises sont extraites de ces sources, puis combinées sous une forme cohérente et unifiée. Le résultat est livré au client sous une forme cohérente et exploitable.

a) *Description Intégration des données*

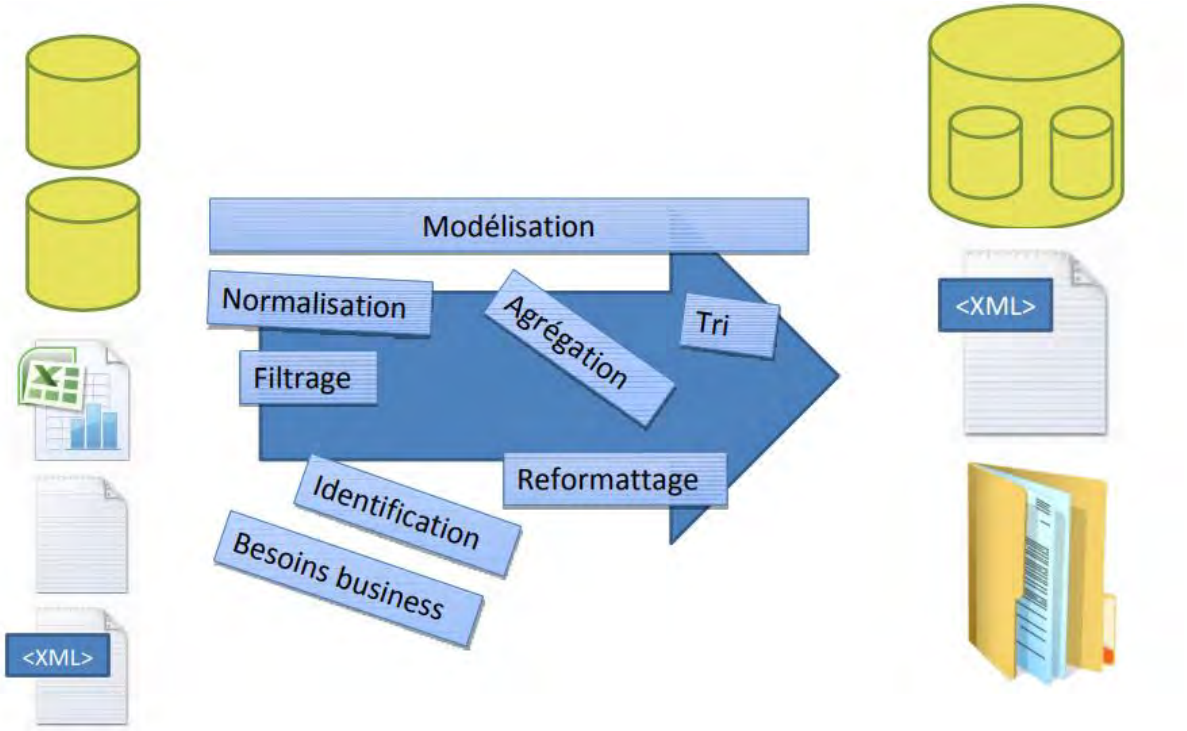

Figure 0-1 : Description 1 intégration données

b) *Schéma Intégration des données*

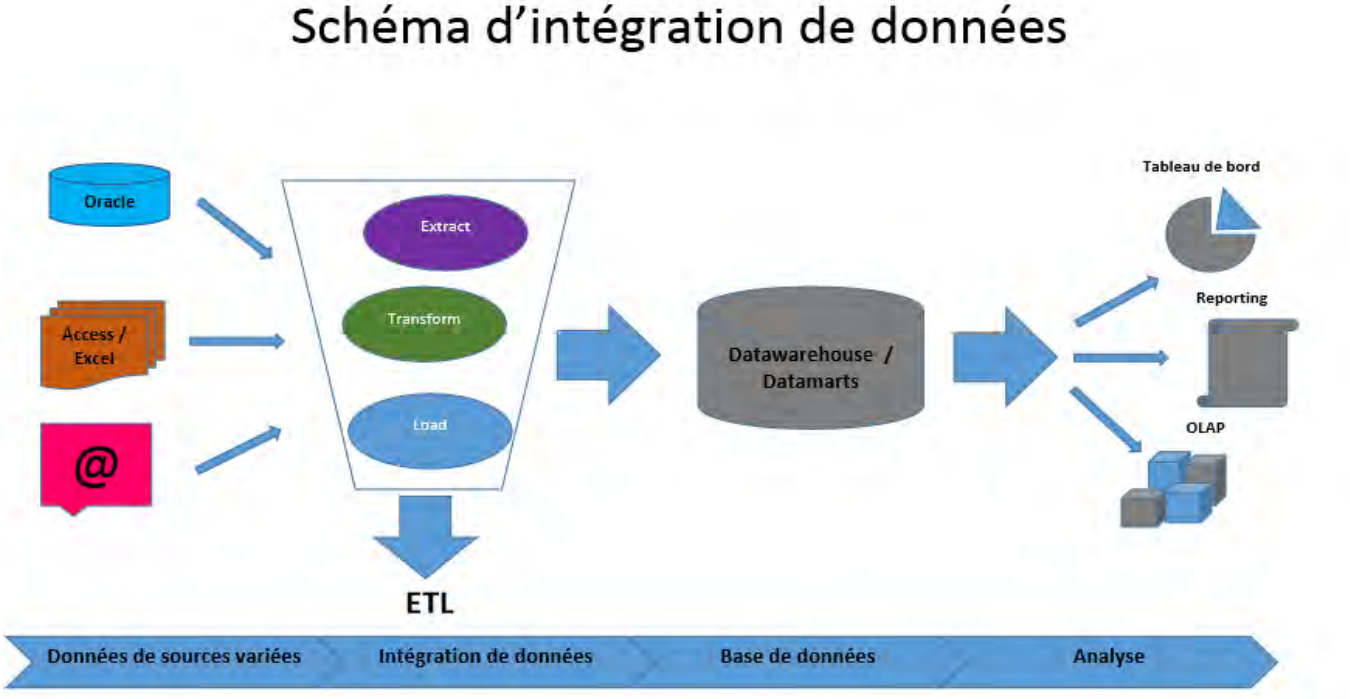

Figure 0-2 : Description 2 intégration données

II. Pourquoi l'intégration de données

Il est possible qu'une entreprise reçoive toutes les données dont elle a besoin, mais ces données sont généralement dispersées dans différentes sources. Par exemple, pour le cas d'usage d'une vue client à 360°, les données qui doivent être combinées peuvent provenir des sources suivantes : système CRM, trafic Web, logiciels utilisés pour les opérations marketing, applications orientées client, systèmes de vente et de succès des clients, partenaires (liste non exhaustive !). Les données provenant de ces différentes sources doivent souvent être rassemblées pour des besoins analytiques ou opérationnels, et il sera parfois difficile pour les ingénieurs de données ou les développeurs de les rassembler efficacement.

Examinons un cas d'usage pour besoins d'analyse. En l'absence de données unifiées, la génération d'un rapport nécessite une analyse, qui implique elle-même de nombreuses opérations : connexion à plusieurs comptes (souvent sur plusieurs sites), recherche des données dans leurs applications natives, copie des données requises, reformatage et nettoyage de ces données.

Pour exécuter cette séquence d'opérations avec le maximum d'efficacité, il est impératif de s'appuyer sur des pratiques d'intégration des données clairement définies et appliquées. Ce cas d'usage présente également les principaux avantages d'une approche d'intégration des données bien pensée :

1. L'intégration des données améliore l'unification des systèmes et la collaboration globale

Les employés de tous les départements, qui se trouvent parfois dans des lieux physiques distants, ont de plus en plus besoin d'accéder aux données de l'entreprise pour des projets individuels ou partagés. Pour leur faciliter la tâche, le département IT doit définir une solution sécurisée pour proposer un accès en libre-service aux données à tous les départements.

Dans la plupart des départements, les employés génèrent, améliorent et enrichissent des données dont le reste de l'entreprise pourrait profiter. L'intégration des données doit donc être une démarche unifiée et collaborative.

2. L'intégration des données fait gagner du temps

Lorsqu'une entreprise prend des initiatives pour l'intégration correcte de ses données, elle réduit considérablement le temps nécessaire à leur préparation et leur analyse. L'automatisation des vues unifiées élimine les tâches manuelles de collecte des données, et les employés n'ont plus besoin d'établir des relations de A à Z lorsqu'ils ont besoin de générer un rapport ou de créer une application.

Par ailleurs, l'utilisation d'outils adaptés à la place du codage manuel fait gagner encore plus de temps à l'équipe de développement (et permet généralement d'économiser des ressources).

Le temps gagné sur ces tâches manuelles peut être utilisé à d'autres fins, par exemple en consacrant plus d'heures à l'analyse et l'exécution pour rendre l'entreprise plus productive et plus compétitive.

3. L'intégration des données réduit les erreurs (et les besoins de modifications)

La gestion des ressources de données d'une entreprise exige un certain nombre de tâches. Pour rassembler les données manuellement et s'assurer que leurs modifications seront complètes et précises, les employés doivent connaître tous les emplacements et tous les comptes qu'ils pourraient avoir à explorer et installer tous les logiciels nécessaires avant de lancer leurs recherches. Si un référentiel de données est ajouté et que tel ou tel employé n'en a pas été informé, son dataset sera incomplet.

En l'absence d'une solution d'intégration qui synchronise les données, les rapports doivent être révisés périodiquement pour tenir compte des changements récents. Avec les mises à jour automatisées des outils d'intégration, les rapports peuvent être générés plus facilement, en temps réel et au moment précis où les utilisateurs en ont besoin.

4. L'intégration des données augmente la valeur des données disponibles

Sur la durée, les efforts d'intégration de données apportent un avantage complémentaire : ils améliorent la valeur des données de l'entreprise. L'intégration de données dans un système centralisé permet de cerner les problèmes de qualité et d'effectuer les améliorations nécessaires, ce qui

permet d'obtenir des données plus précises, ce qui est le fondement même des analyses de qualité.

III. L'ÉVOLUTION DE L'INTÉGRATION DES DONNÉES

1. Data warehouse

Ce sont des « entrepôts de données », nés du besoin de centraliser les données des différentes bases opérationnelles de l'entreprise. Ceci dans le but d'effectuer des analyses sur plusieurs dimensions. Ils présentent les contraintes et limitations suivantes:

- ✓ Il est difficile de concevoir un schéma unique pour l'intégration de toutes les données provenant de diverses sources,
- ✓ Il est difficile de transformer des données avec les outils ETL,
- ✓ Les mises à jour des Data Warehouses sont très lentes car elles sont en général effectuées par des batchs journaliers, hebdomadaires ou même mensuels.

2. Data lake

Concept popularisé par l'émergence des plateformes Hadoop. Les Data Lakes, ou lac de données, permettent de s'affranchir des problèmes de modélisation de schéma d'intégration unique en permettant d'insérer toutes les données, quelle que soient leur nature et leur origine. Ils simplifient ainsi les processus d'alimentation traditionnels mais présentent les contraintes et limitations suivantes :

- ✓ L'exploitation du data lake demeure en mode batch. Même si les batch peuvent désormais être lancés à tout moment, cela reste en deçà du temps réel où toute nouvelle information serait immédiatement prise en compte dès sa publication,
- ✓ Le principe du data lake est que les données sont déversées sans transformation préalable, laissant ainsi cette tâche aux consommateurs. La contrainte qui est levée pour l'écriture est déplacée à la lecture, schema-on-write et schema-on-read.

3. Le Streaming data processing

C'est le traitement des données d'un flux au fur et à mesure qu'elles arrivent. Il présente l'avantage de permettre aux utilisateurs et aux systèmes intéressés par le traitement du flux, de pouvoir réagir instantanément aux évolutions en cours. Ce que ne permettent pas les systèmes d'intégration vus précédemment. C'est un paradigme d'intégration qui a une application très intéressante dans les architectures à base de micro services.

4. L'impact de l'avénement streaming data processing sur les métiers autour du traitement des données

La transition du Data Warehouse vers le Data Lake et le streaming se fait aussi avec un changement des rôles des différents profils intervenants sur la chaine d'acquisition et d'exploitation des données.

Avec le Data Warehouse, des ingénieurs spécialisés avec les outils ETL ont le principal rôle dans le processus d'acquisition des données. Mais avec les Data Lake et le streaming, ce sont les ingénieurs applicatifs qui ont la responsabilité de publier directement leurs données à destination des autres consommateurs. Même si la publication des données ne se fait plus avec les logiciels ETL spécialisés, il n'en demeure pas moins que les développeurs doivent toujours gérer l'échange de données entre différentes applications utilisant différents formats et assurer la cohérence des données partagées entre ces applications. Ce qui peut être finalement vu comme un processus ETL et décris l'évolution de l'intégration des données en entreprise vers le streaming data processing pour pouvoir réagir toujours plus vite à l'arrivée de données provenant de sources multiples.

IV. Intégration de données dans les entreprises performantes

L'intégration de données ne se présente pas sous la forme d'une solution universelle ou standard : la formule à appliquer peut varier en fonction des différents besoins de l'entreprise. Voici quelques cas d'usage courants pour les outils d'intégration des données :

1. Exploiter les big data

Les data lakes sont souvent très volumineux et très complexes. Des sociétés telles que Google ou Facebook traitent un flux continu de données provenant de plusieurs milliards d'utilisateurs. Ce niveau de consommation de l'information est communément désigné par le terme « big data ». Plus le nombre de grandes sociétés impliquées dans le traitement des big data augmente, plus les entreprises ont accès à des volumes de données importants dont elles peuvent envisager l'exploitation. C'est pour cette raison qu'il est impératif pour la plupart des entreprises de déployer des efforts poussés pour l'intégration de leurs données.

2. Créer un data warehouse

Les initiatives d'intégration des données – en particulier dans les grandes sociétés – ont souvent pour but de créer des data warehouses qui combinent plusieurs sources de données dans une seule base de données relationnelles. Les data warehouses permettent aux utilisateurs de formuler des requêtes, compiler des rapports, produire des analyses et récupérer des données sous un format cohérent.

3. Simplifier la Business Intelligence

En fournissant une vue unifiée des données provenant de nombreuses sources, l'intégration de données simplifie les processus de Business Intelligence (BI). Les entreprises peuvent visualiser plus facilement et comprendre plus rapidement les datasets disponibles et récupérer ainsi des informations exploitables sur l'état de l'entreprise. Avec l'intégration des données, les analystes peuvent compiler plus d'informations et obtenir des évaluations plus précises sans être submergés par des volumes de données élevés.

Contrairement à l'analytique, la BI n'utilise pas l'analyse prédictive pour faire des projections : elle se concentre plutôt sur la description du présent et du passé pour faciliter les prises de décision stratégiques. Cette utilisation de l'intégration de données est bien adaptée au data warehouse, dans lequel les informations de vue d'ensemble sous un format facile à consommer s'alignent parfaitement.

4. Opérations ETL et intégration des données

Les opérations d'extraction, de transformation et de chargement (ETL) forment un processus d'intégration à part entière dans lequel les données sont extraites du système source et livrées au data warehouse. Il s'agit d'un processus continu que le data warehousing exécute pour transformer plusieurs sources de données en informations cohérentes et utiles destinées à la Business Intelligence et aux analyses.

V. Défis de l'intégration des données

Regrouper les données de plusieurs sources et les transformer en un dataset unifié et stocké dans une structure unique est un défi technique en soi. Plus les entreprises développent des solutions d'intégration de données, plus elles sont obligées de créer des processus spécifiques pour déplacer sans cesse les données là où les utilisateurs en ont besoin. Bien que ces opérations permettent de réaliser des économies de temps et d'argent à court terme, leur implémentation peut être gênée par de nombreux obstacles.
Voici quelques défis courants auxquels les entreprises font face lors de la construction de leurs systèmes d'intégration :

1. Défis, objectifs et solutions

Les entreprises savent généralement ce qu'elles attendent de l'intégration de données : la solution à un défi bien précis. Ce à quoi elles ne pensent pas toujours, au chemin à prendre pour y parvenir. L'entreprise qui décide d'implémenter une solution d'intégration des données doit identifier les types de données à collecter et analyser, les sources de ces données, les systèmes qui vont les utiliser, les types d'analyse à effectuer et la fréquence à laquelle les données et rapports devront être mis à jour.

2. Données des systèmes legacy

Les efforts d'intégration peuvent exiger l'inclusion des données stockées dans les systèmes legacy. Noter toutefois que ces données legacy ne sont pas

toujours associées à des marqueurs tels que les heures et dates des activités (ces marqueurs sont généralement prévus dans les systèmes plus modernes).

3. Données liées à de nouvelles exigences internes

Les systèmes plus récents génèrent des types de données différents (par exemple, données non structurées ou données en temps réel) et à partir de sources variées telles que les vidéos, les objets IoT, les capteurs et le cloud. Trouver la solution qui permettra d'adapter rapidement votre infrastructure d'intégration pour répondre aux nouvelles exigences de ces données devient crucial pour le succès de votre entreprise, mais aussi extrêmement difficile dans la mesure où le volume, la vitesse, les nouveaux formats de ces données font apparaître de nouveaux défis.

4. Données externes

Les données provenant de sources externes peuvent présenter un niveau de détail moins élevé que les données provenant de sources internes, ce qui ne permet pas toujours de les examiner avec la même rigueur. De plus, les termes des contrats signés avec certains fournisseurs externes peuvent rendre difficile le partage des données à l'échelle de l'entreprise.

5. Maintenir le rythme

Lorsque le système d'intégration des données est en place et opérationnel, la tâche n'est pas terminée. En effet, il est de la responsabilité de l'équipe Données de maintenir les efforts d'intégration au niveau des meilleures pratiques du secteur tout en respectant les exigences les plus récentes de l'entreprise et des organismes de réglementation.

VI. Comment intégrer les données de l'entreprise

Il existe plusieurs façons d'intégrer les données en fonction de la taille de l'entreprise, des besoins à satisfaire et des ressources disponibles :

1. L'intégration manuelle des données

Elle n'est ni plus ni moins que le processus par lequel un utilisateur donné collecte manuellement les données nécessaires à partir de différentes sources (en accédant directement à leurs interfaces), puis les nettoie selon les besoins et les regroupe dans un data warehouse. Cette méthode est peu efficace et peu cohérente, et elle est déconseillée pour la plupart des entreprises, sauf sans doute les plus petites, dont les ressources de données sont très réduites.

2. L'intégration de données par middleware

Elle est une approche dans laquelle une application middleware agit en tant que médiateur et aide les utilisateurs à normaliser les données et à les injecter dans le pool de données de référence. (Pensez aux anciens équipements électroniques dont les connecteurs sont obsolètes mais qui sont prêts à reprendre vie avec un simple adaptateur.) Les applications legacy fonctionnent rarement bien avec les applications les plus récentes : le rôle du middleware est d'intervenir chaque fois qu'un système d'intégration de données ne parvient pas à accéder par lui-même aux données d'une application legacy.

3. L'intégration basée sur les applications

Elle est une approche dans laquelle les applications identifient, récupèrent et intègrent les données requises. Le logiciel d'intégration doit assurer la compatibilité des données provenant de différents systèmes pour pouvoir les transmettre d'une source à une autre.

4. L'intégration avec accès uniforme

Elle se concentre sur la création d'un front-end (interface) qui présente sous une forme cohérente les données accessibles à partir de différentes sources (en réalité, les données sont conservées dans leur source originale). Avec cette méthode, des systèmes de gestion de base de données orientés objets peuvent être utilisés pour créer une apparence d'uniformité entre des bases de données différentes.

5. L'intégration avec stockage commun

Elle est l'approche la plus fréquemment utilisée pour le stockage des données résultant d'opérations d'intégration. Une copie des données de la source originale est conservée dans le système intégré puis traitée pour une vue unifiée. Cette solution est donc différente de l'intégration avec accès uniforme, qui conserve les données dans leur source respective. L'approche avec stockage commun est le principe sous-jacent de la solution traditionnelle du data warehousing.

6. Caractéristiques à rechercher dans un outil d'intégration des données

a) Un grand nombre de connecteurs

La diversité des systèmes et applications étant considérable, plus votre outil d'intégration des données disposera de connecteurs intégrés, plus vos équipes gagneront du temps.

b) Open source

Les architectures open source offrent généralement une plus grande flexibilité tout en facilitant la prévention du provisionnement captif.

c) La portabilité

Elle est importante, au moment où les entreprises adoptent de plus en plus les modèles de cloud hybride, pour construire vos intégrations de données une seule fois et les exécuter partout.

d) Facilité d'utilisation

Les outils d'intégration des données doivent être faciles à apprendre et utiliser, et proposer une interface utilisateur graphique pour faciliter la visualisation des pipelines de données.

e) Modèle de tarification transparent

Votre fournisseur d'outils d'intégration de données ne devrait pas vous demander d'augmenter le nombre de connecteurs ou les volumes de données.

7. Complexité sur l'intégration

a) Data profiling

Bien connaître le type de données à traiter afin de prévoir tous les cas possibles.

b) Monitoring

Surveiller les données sources Récolter des méta-données à chaque étape, détection précise d'erreurs et possibilité de rollback.

c) Éliminer les goulets d'étranglement

- ✓ Manque de mémoire vive
- ✓ Opérations inefficaces dans la DB
- ✓ Trop d'opérations d'entrée/sortie
- ✓ Reconstruction d'agrégats inefficace
- ✓ Écritures inutiles suivies de lecture
- ✓ Filtrage trop tardif des données

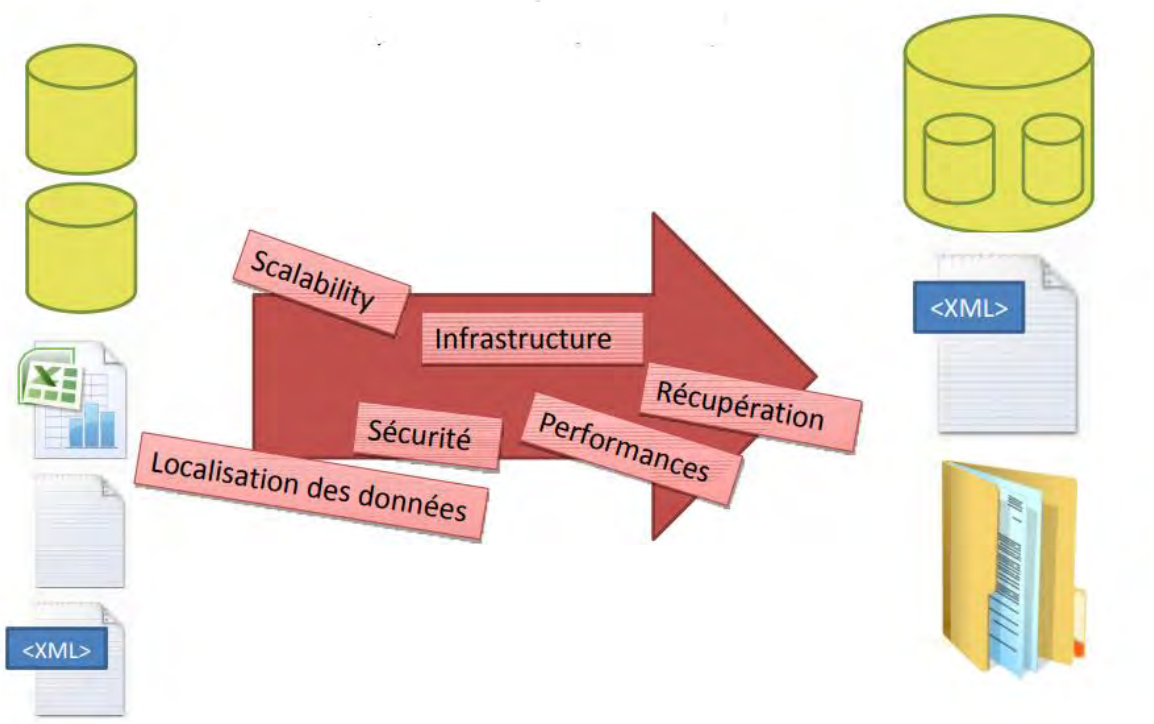

Figure 0-3 : Complexité sur l'intégration de donnée