Bloor

Livre blanc

Livre blanc de Bloor

Auteur : **Philip Howard**

Date de publication **Janvier 2017**

La gestion de Data Lake

« La gestion de Data Lake doit favoriser la productivité et la collaboration, tout en simplifiant et en accélérant l'identification de données fiables et l'accès à celles-ci au sein du Data Lake. »

Auteur **Philip Howard**

Synthèse

Ce livre blanc met l'accent sur les défis de gestion qui se présentent lorsque vous implémentez un Data Lake, et les coûts qu'impliquent les différentes approches de la gestion de Data Lake. Même si ce rapport a été commandité par Informatica et Cognizant, il traite de problèmes génériques. Les chiffres et les captures d'écran utilisés pour illustrer nos propos sont fournis par Informatica, de même que les différentes citations d'utilisateurs incluses. Néanmoins, tous les autres aspects de nos discussions ne dépendent d'aucun fournisseur de technologies ou de solutions particulier.

Avant de commencer, nous devons définir clairement la terminologie. La définition d'un Data Lake fait l'objet d'une homogénéité certaine (mais pas totale) au sein du secteur. Le plus important semble être la distinction entre Data Lake d'un côté et data warehouse (ou Data mart) de l'autre. À la différence du premier, ce dernier présente des schémas définis pour répondre à des exigences de reporting et d'analyse connues. En revanche, les Data Lakes concernent davantage le stockage de données à des fins d'exploration et d'organisation. Certains estiment également qu'un Data Lake ne se limite pas nécessairement à un seul référentiel de données. Néanmoins, pour les besoins de ce livre blanc, nous considérerons qu'un Data Lake est un référentiel évolutif qui ne vous oblige pas à déclarer un schéma lors de l'ingestion des données, et qui prend en charge l'exploration et l'organisation de celles-ci.

On peut aussi se demander en quoi la gestion de Data Lake est essentielle pour tirer profit de votre Data Lake. C'est parce que les indicateurs de réussite des Data Lakes s'articulent fondamentalement autour de la vitesse et de la fiabilité. Les entreprises ont besoin de données dans des délais spécifiques et elles ont besoin que ces données respectent des niveaux de fiabilité spécifiques. En d'autres termes, la gestion de Data Lake doit favoriser la productivité et la collaboration, tout en simplifiant et en accélérant l'identification de données fiables et l'accès à celles-ci au sein du Data Lake.

Ce sont ces considérations relatives à la gestion de Data Lake que nous aborderons dans le présent livre blanc, qui est subdivisé en deux parties. Dans la première, nous discuterons des défis que pose la gestion d'un Data Lake, et nous aborderons les différentes approches possibles pour gérer un Data Lake, principalement en nous penchant sur le débat « plate-forme de pointe ou intégrée ». Dans la seconde, nous examinerons les répercussions des différentes approches de gestion de Data Lake sur le coût total de possession (TCO).

« **Un Data Lake est un référentiel évolutif qui ne vous oblige pas à déclarer un schéma lors de l'ingestion des données, et qui prend en charge l'exploration et l'organisation de celles-ci.** »

Exigences liées à la gestion de Data Lake

La gestion de Data Lake entraîne différents problèmes qui vont de la manière dont vous y placez vos données à celle dont vous procédez aux analyses. La **figure 1** illustre le grand nombre d'éléments potentiellement impliqués. Même si nous n'abordons pas chacun des éléments indiqués, nous analyserons les principales préoccupations une à une.

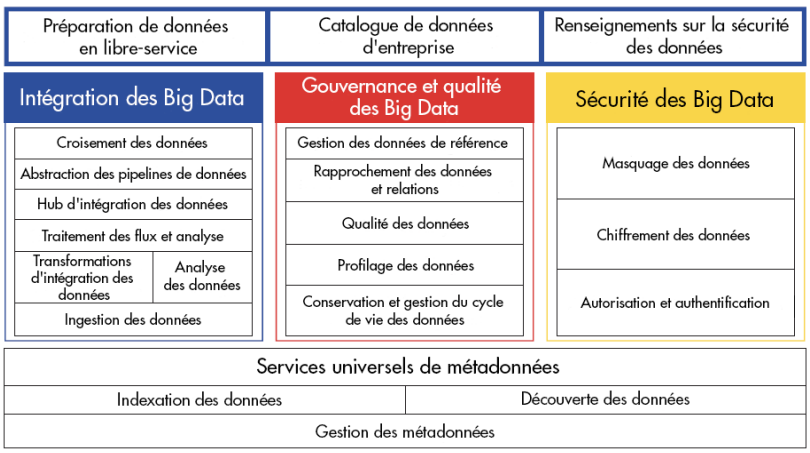
« Le catalogage désigne la découverte des données placées dans votre Data Lake, ainsi que la création d'un catalogue de toutes les ressources de données qui s'y trouvent et dans lequel les utilisateurs peuvent effectuer des recherches pour trouver des informations pertinentes. »

Ingestion et transformation des données

La première question concerne les efforts nécessaires pour placer les données dans votre Data Lake. Dans la pratique, cela recouvre plusieurs questions distinctes : le déplacement physique des données, leur analyse au format que requiert le Data Lake, puis leur transformation pour permettre à vos utilisateurs d'utiliser vos ressources de données.

Quelle que soit la méthode de chargement (c'est-à-dire le déplacement physique des données requises), la question centrale concerne les efforts et, par conséquent, le temps requis pour réaliser cette tâche. Se pose également la question des efforts (manuels ou automatisés) nécessaires pour atteindre l'objectif ultime, c'est-à-dire analyser et transformer des données brutes en informations adaptées aux objectifs.

Figure 1 : éléments à prendre en compte pour la gestion de Data Lake



Catalogage des données

Différentes tâches sont nécessaires pour que les données de votre Data Lake soient utilisables et compréhensibles. Il s'agit notamment du catalogage et de la préparation des données. Nous pensons qu'il s'agit de l'étape suivante à prendre en compte après l'ingestion, même si nous sommes conscients que certaines entreprises optent pour un ordre différent.

Le catalogage désigne la découverte des données placées dans votre Data Lake, ainsi que la création d'un catalogue de toutes les ressources de données qui s'y trouvent et dans lequel les utilisateurs peuvent effectuer des recherches pour trouver des informations pertinentes. Cette fonctionnalité, dont un exemple est indiqué dans la **figure 2**, est indispensable pour permettre aux utilisateurs de découvrir facilement les informations accessibles dans le Data Lake. Certaines entreprises ont tenté de développer leurs propres outils de catalogage mais nous ne recommandons pas cette approche : il y a trop de données et de métadonnées, celles-ci changent trop fréquemment et, dans tous les cas, les processus personnalisés sont coûteux, propices aux erreurs et difficiles à gérer.

Il faut débarrasser le catalogage des longs processus manuels pour permettre la préparation en libre-service et à grande échelle pour les consommateurs de données. Il est donc nécessaire de disposer d'une fonctionnalité de catalogage automatisée, capable d'extraire le sens des données sans intervention humaine explicite. Les métadonnées (techniques, commerciales et opérationnelles) sont indispensables à la découverte des ressources de données. Vous devez fournir une indexation automatique et des fonctionnalités de recherche (sémantique). Cela sera clairement avantageux si les produits de catalogage peuvent tirer automatiquement parti des métadonnées potentiellement capturées pendant le processus d'ingestion. L'apprentissage machine et les capacités similaires seront également utiles pour déduire la structure des ressources de données et les relations entre ces ressources. De plus, l'apprentissage machine contribuera à garantir l'automatisation du processus de catalogage.

Enfin, il sera utile si le catalogue contient des informations détaillées sur des données provenant de sources extérieures au Data Lake. Par exemple, si les données d'un data warehouse ne doivent pas être répliquées dans le Data Lake, il conviendra alors de cataloguer également les données du data warehouse.

Sécurité des données

Les Data Lakes contiennent fréquemment des données sensibles. Dans la plupart des cas, les analystes métiers et les spécialistes des données qui accèdent à votre Data Lake ne seront pas autorisés à voir ces informations sensibles. Il faut donc identifier (généralement à l'aide d'un outil de profilage des données) et masquer ces données avant qu'elles ne soient mises à la disposition des analystes. Il est raisonnable de considérer les spécialistes des données et les analystes métiers comme des développeurs : ils développent simplement des analyses plutôt que des applications, et les mêmes principes s'appliquent aux uns comme aux autres du point de vue des informations sensibles. Et si l'on considère que le Règlement général sur la protection des données (RGPD) de l'UE prévoit une amende pouvant s'élever à 4 % du chiffre d'affaires en cas d'atteinte à la confidentialité des données, on aurait tort d'ignorer la nécessité de masquer les données avant leur préparation. Il serait tout aussi insensé de faire reposer ce masquage sur des processus exigeant beaucoup de travail, potentiellement incapables de fournir rapidement ou systématiquement des données sécurisées aux utilisateurs.

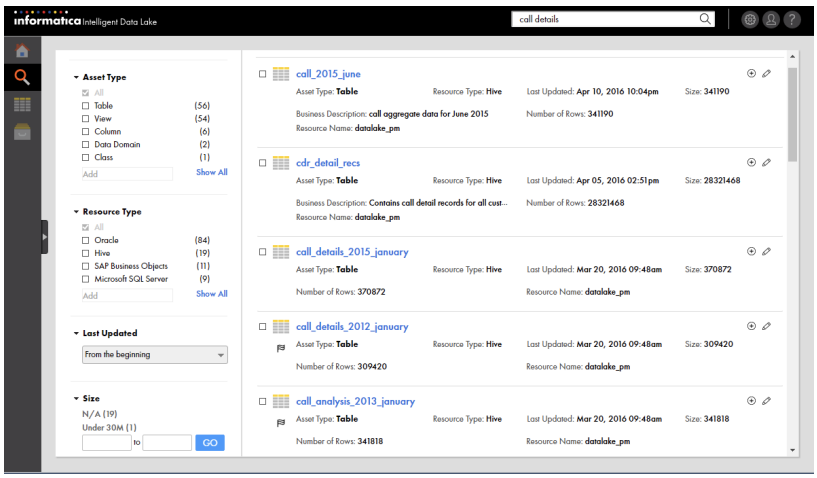
Voici certaines des capacités requises :

- **Profilage des données.** Vous devrez pouvoir réaliser un profilage autonome des données pour découvrir quelles sont les données sensibles à masquer et protéger. Notons que certains fournisseurs de solutions de masquage des données proposent des capacités de profilage spécifiquement conçues pour découvrir les données sensibles, et non des capacités de profilage plus génériques.
- **Masquage des données.** Différentes entreprises proposent des produits autonomes fournissant un masquage statique des données, ce qui offre une protection au sein du Data Lake. Néanmoins, il ne s'agit pas uniquement de masquer les données dans le Data Lake. Les données d'origine externe devront être profilées (et masquées) lorsqu'elles arrivent dans le Data Lake, mais les données générées en interne doivent être masquées avant d'arriver dans le Data Lake (après tout, les données des systèmes opérationnels sont tout aussi sensibles que celles de votre Data Lake). Mais cela ne se passe pas toujours de cette manière. Par conséquent, le profilage des données doit être opérationnel tant dans les environnements conventionnels que dans le Data Lake, et il en va de même pour le masquage.

Il est également important de pouvoir procéder à des vérifications (audits) pour s'assurer que les données ont été masquées, et de pouvoir prouver que les données masquées ne peuvent pas faire l'objet d'une ingénierie inverse. Cela signifie qu'on ne peut pas considérer le chiffrement comme suffisant. Comme le masquage fait partie intégrante de votre solution de sécurité des données, en plus de répondre aux exigences de conformité, il sera utile que l'audit de votre masquage s'intègre avec votre environnement de sécurité au sens large.

> **« Nos données clients étaient souvent obsolètes, inexactes ou incomplètes. »**
>
> **Barbara,**
> Chief Data Governance Officer

Figure 2 : résultats de recherche dans un catalogue de données



> **La préparation et le nettoyage des données nous prenaient 2-3 semaines.**
>
> **Vishal,**
> VP Data Architecture

Qualité des données

La qualité des données pose deux questions majeures : « comment » l'obtenir et « où » l'obtenir. Pour la question du « comment », il vous faut un maximum d'automatisation. Cela est dû au fait qu'à chaque fois que vous devez intervenir manuellement au niveau d'un processus de qualité des données (ou de tout processus), vous a) introduisez un risque d'erreur et b) ralentissez ce processus. Les deux font grimper vos coûts. Par conséquent, vous devez automatiser au maximum vos processus de qualité des données. Cela nécessite généralement une approche basée sur les règles, éventuellement avec l'intégration de l'apprentissage machine. Mais ce qui compte le plus, c'est de le faire, la manière compte moins.

La deuxième question concerne l'endroit où les processus de qualité interviennent, et cela conduit à une troisième question liée aux outils utilisés pour garantir la qualité des données. Premièrement, les opérations de nettoyage et de déduplication des données doivent-elles être réalisées avant ou après le placement des données dans le Data Lake ? Utilisez-vous des processus de qualité des données traditionnels dans votre environnement opérationnel, ou bien les capacités de nettoyage de votre produit de préparation des données (voir ci-dessous) sont-elles suffisantes ? Si les données proviennent de sources externes, la réponse est clairement « après », mais dans le cas contraire, ce n'est pas aussi évident. Pour les données de production, nous sommes davis que, comme la qualité des données est aussi importante à des fins opérationnelles qu'à des fins d'analyse, elles doivent être pré-nettoyées. Il en va de même pour les données sensibles : il n'y a pas vraiment d'intérêt à masquer des données incorrectes.

C'est plus complexe pour les données en cours de transfert dans le Data Lake. Lesflux comportent souvent des événements manqués, des événements qui n'arrivent pas par ordre chronologique et même des événements en double (imaginez par exemple plusieurs antennes-relais qui collectent des données sur le même appel téléphonique). Idéalement, vous disposez d'une plate-forme de traitement des flux qui dispose de fonctionnalités intégrées pour traiter de telles anomalies, mais si ce n'est pas le cas, vous devrez alors traiter le nettoyage après le transfert des données. C'est quelque chose d'important et cela sous-entend que vous aurez besoin de processus de qualité des données formels capables de s'exécuter dans votre environnement de Data Lake.

Trois situations nécessiteront donc le nettoyage et la déduplication des données : dans les environnements opérationnels avant de déplacer les données dans votre Data Lake, dans le Data Lake et, enfin, dans le cadre des processus de qualité des données que prend en charge votre outil de préparation des données. Comme votre objectif est de vous assurer que la qualité des données est mise en œuvre de manière exhaustive dans l'ensemble des projets de l'entreprise, il est probablement absurde d'avoir trois ensembles de produits totalement différents, compte tenu du coût des formations et des licences nécessaires, et du manque de transfert de compétences entre de tels produits. S'il peut être intéressant d'envisager des outils de préparation des données ciblant les utilisateurs métiers, nous pensons qu'en matière de qualité des données traditionnelle, il est essentiel qu'un seul ensemble d'outils soit disponible pour traiter à la fois les environnements traditionnels et le Data Lake.

Préparation des données

La préparation des données avant l'analyse implique plusieurs processus différents. Les outils de préparation des données doivent tout d'abord tirer parti des catalogues, de la recherche dans les catalogues et des métadonnées qui ont déjà été décrites. Ensuite, ils nécessitent des fonctions de profilage et de nettoyage en libre-service des données, comme nous l'avons décrit ci-dessus, mais d'une manière adaptée aux analystes métiers, par opposition aux gestionnaires de données. C'est important, car les analystes devront prendre des décisions concernant les données incomplètes (par exemple les valeurs nulles), les données incohérentes (comme les codes non standards) et les problèmes associés.

Néanmoins, la préparation des données implique également le croisement des données (parfois appelé « unification » lorsqu'il est réalisé à grande échelle et, souvent, lorsqu'il concerne des données non présentes dans le Data Lake). Cela

implique essentiellement trois processus : l'ingestion des données non stockées dans votre Data Lake, l'obtention des données dans un format cohérent et la jonction des données. Ce sont des processus importants qui peuvent nécessiter des fonctionnalités supplémentaires.

Lorsque des données externes issues d'une base de données doivent être consommées, il est possible d'utiliser des connecteurs adéquats (idéalement, plutôt des connecteurs natifs que des interfaces ODBC ou JDBC). Néanmoins, il est fréquent que les données externes proviennent d'environnements applicatifs, qui peuvent être ou non basés dans le Cloud. Dans tous les cas, cela implique la prise en charge d'interfaces de programmation d'applications (API) adéquates pour se connecter à Salesforce.com, Workday, SAP et autres environnements applicatifs. Le nombre écrasant d'applications, disposant chacune de leurs propres API, peut alors nécessiter une fonctionnalité de gestion des API.

La cohérence des formats n'est pas une question simple. Par exemple, si vous souhaitez combiner des données textuelles et relationnelles, vous devrez analyser le texte en premier pour en extraire des termes significatifs qui pourront être mis en correspondance avec ceux que contiennent vos tables relationnelles. Réaliser ce processus manuellement revient à analyser les éléments communs tels que les prénoms et les adresses e-mail. Si vous joignez un grand nombre d'ensembles de données, ce processus peut s'avérer très fastidieux. Les outils de préparation des données capables de reconnaître les clés de jointure potentielles et de vous les recommander (entre autres actions) vous feront considérablement gagner en productivité. L'exemple proposé dans la **figure 3** illustre cela, avec des recommandations réalisées par le logiciel.

Enfin, même si la gouvernance est largement cachée à l'utilisateur, vous devez l'intégrer, de même que différentes fonctionnalités qui optimisent productivité et collaboration. Plus particulièrement, les fonctionnalités de workflow qui capturent les processus de préparation des données à des fins de réutilisation et de partage sont fondamentales.

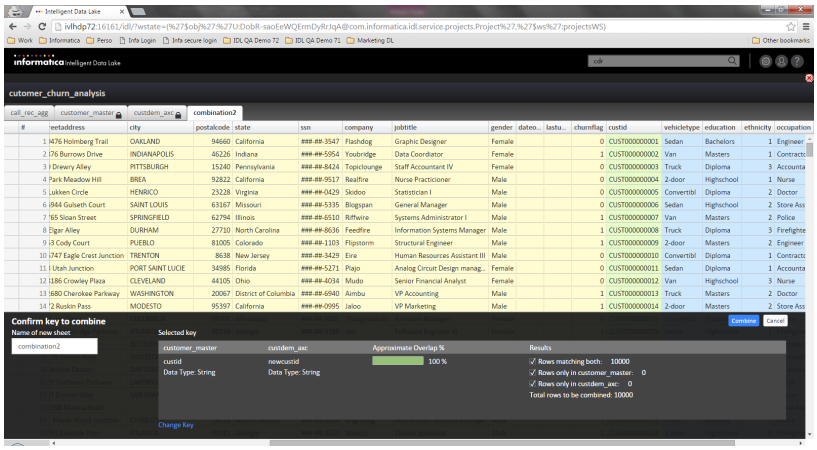
Gouvernance des données

La gouvernance et la conformité vont de pair avec la sécurité. Dans le contexte d'un Data Lake, vous devez appliquer la gouvernance et la conformité en douceur. Après tout, il s'agit d'un environnement en libre-service. Il convient néanmoins de respecter certaines exigences métiers et réglementaires. Les fonctionnalités de gouvernance doivent donc être intégrées à l'environnement de Data Lake pour permettre au département informatique de surveiller qui accède à quelles données, comment celles-ci sont combinées et dans quel but elles sont utilisées. En plus de satisfaire aux exigences de conformité, les fonctionnalités de gouvernance peuvent également optimiser la collaboration.

Mentionnons également la gestion des données de référence (MDM), que l'on associe souvent à la gouvernance des données. Le fait est que l'on utilise souvent les Data Lakes pour compléter les systèmes MDM et de gestion de la relation client (CRM), et pour fournir une vision globale et élargie des clients, des produits, des fournisseurs, etc. L'idée, c'est que les données relationnelles traditionnelles restent dans vos systèmes existants, mais que les informations supplémentaires comme les informations de contact du centre d'appels, les commentaires sur les médias sociaux, les profils LinkedIn et autres données de ce type soient conservées dans le Data Lake. De plus, le traitement évolutif qu'offrent les plates-formes de données comme Hadoop

> **« Nécessité de regrouper les silos d'informations et de centraliser les données. »**
>
> **Hamilton,**
> Head of
> Global Data Strategy

Figure 3 : préparation des données avec recommandations



peut permettre d'effectuer très rapidement le rapprochement des données, la découverte de relations (par exemple maison-propriété) et les liaisons d'enregistrements sur de très grands ensembles de données. Cela vous permet d'obtenir une vision d'ensemble de vos clients, produits, patients, fournisseurs et autres domaines d'entité.

Accès aux données et consommation des données

Enfin, les utilisateurs veulent pouvoir exécuter des requêtes et élaborer des analyses au sein de votre Data Lake. Celui-ci doit donc offrir un accès efficace aux données, via des connecteurs et des modèles publication-abonnement prédéfinis. Pour y parvenir, on a souvent recours à la préparation des données, au catalogage des données ou aux outils de type « hub d'intégration de données » qui se connectent directement à ces environnements tiers. Dans ce contexte, la latence et la modalité sont toutes deux importantes, donc vous devez opter pour des connecteurs prédéfinis, des options « push/pull » et des mécanismes de publication-abonnement automatisés.

Synthèse des exigences

Il apparaît clairement que la gestion d'un Data Lake présente de nombreuses variables dont la gestion est essentielle pour tirer profit des Big Data. Cela nous place devant un dilemme évident lors de l'évaluation des différentes approches de la gestion de Data Lake. Plus les fonctionnalités de gestion de Data Lake sont incohérentes et cloisonnées entre elles, plus il est difficile de justifier une approche centrée sur les performances. Combien cela coûte-t-il d'avoir six ou sept solutions différentes qui imposent de longs processus manuels ? Nous allons voir cela dans la prochaine section, car il est important de faire le bilan des choix disponibles. Toutefois, avant d'entamer la discussion sur le coût total de possession, nous avons quelques remarques préliminaires.

Tout d'abord, nous pensons que le codage manuel ne peut absolument pas être un bon choix pour l'une quelconque des fonctions dont nous avons parlé. Si nous prenons l'exemple de l'intégration des données, un domaine dans lequel les méthodes manuelles restent bien trop courantes, Bloor Research a réalisé de nombreuses études sur les coûts d'utilisation des différentes approches de l'intégration des données, avec toujours la même conclusion : tous comptes faits, le codage manuel constitue une fausse économie. Pour consulter notre dernier rapport sur le sujet, rendez-vous à l'adresse ***http://www.bloorresearch.com/research/white-paper/comparative-costs-and-uses-for-data-integration-platforms-in-agile-enterprises/***. En bref, vous ne devez jamais avoir recours à des méthodes manuelles pour les projets d'infrastructure. Cela comprend les projets partiels reposant sur des produits open source tels que Kafka ou Sqoop. Conclusion : c'est l'automatisation qui permettra d'améliorer la productivité des environnements de Data Lake, et vous ne profiterez pas de cette automatisation sans utiliser une solution basée sur une plate-forme capable de masquer les complexités sous-jacentes de l'environnement.

Mentionnons également l'utilisation des métadonnées. Une approche de la gestion de Data Lake axée sur la réduction du coût total de possession tire grandement parti des métadonnées issues de l'ingestion et de la transformation, de la gouvernance et des outils de qualité des données au catalogage et à la préparation des données. Bien qu'il soit possible d'utiliser des produits d'échange de métadonnées pour essayer d'établir des passerelles entre différents produits, dans la pratique, les coûts d'intégration élevés des différents produits nuisent à l'efficacité économique de cette approche.

Par conséquent, et avant de poursuivre, nous avons deux recommandations claires. Tout d'abord, ignorez toutes les solutions nécessitant un codage manuel, même lorsque vous pouvez tirer parti d'offres open source. Ce type d'approche aboutit à des solutions incohérentes et non réutilisables. De même, ignorez les produits qui ne reposent pas sur les métadonnées ou sont incapables de partager celles-ci. Ils aboutiront à des solutions cloisonnées, si tant est qu'il soit possible d'implémenter une solution complète sans métadonnées.

> **« Nécessité d'accéder à nos données... nous avions des silos de données... cela provoquait également des silos analytiques. »**
>
> **Raman,**
> VP Enterprise Analytics Management

Coût total de possession

Ici, nous abordons les principes régissant le calcul du coût total de possession. Il y a bien entendu des frais de licence et des coûts d'abonnement. Des frais de maintenance peuvent également s'appliquer. Et dans le cas des Data Lakes, il peut y avoir des coûts matériels associés ainsi que d'autres coûts si l'un des logiciels que vous vous proposez d'utiliser ne s'exécute pas sur la même plate-forme que votre Data Lake. En d'autres termes, il peut y avoir des coûts de déploiement supplémentaires, en plus des logiciels.

Comme nous considérons que ces éléments entrent naturellement dans le calcul du coût total de possession, nous n'en parlerons pas davantage. Tous ces éléments doivent faire partie de votre coût total de possession. Nous voulons nous concentrer sur les coûts moins évidents et moins visibles qui sont susceptibles d'être associés aux différentes approches de la gestion de Data Lake. En voici quelques-uns :

- **Coûts de déploiement et de développement.** Dans la gestion des Data Lakes, le développement intervient pendant l'ingestion et la phase ETL (extraire, transformer et charger), et, éventuellement, s'il faut développer des connecteurs et/ou des API en raison de leur absence dans la solution de base. Dans ce dernier cas, il est donc nécessaire de déterminer quelles sont les sources de données à intégrer et quels efforts de déploiement manuels seront nécessaires (le cas échéant) pour développer les capacités adéquates afin de les intégrer. Les entreprises doivent également garder à l'esprit que, comme des sources de données supplémentaires peuvent être ajoutées au fil du temps, la disponibilité de kits de développement logiciel pour développer des adaptateurs personnalisés est essentielle. Concernant toutes les capacités de gestion de Data Lake, quelle sera la quantité de développement manuel nécessaire pour développer les processus adéquats ? Il faut quantifier cela en nombres d'heures passées, puis convertir ce chiffre en coûts. La productivité et la réutilisation seront très importantes ici : si vous avez recours au codage manuel, ou à d'autres méthodes qui ne prennent pas en charge la réutilisation, chaque nouveau développement présentera des coûts qui devront être supportés à nouveau, tandis que la réutilisation permettra de les minimiser. Et il ne faut pas oublier que les personnes font des erreurs et que celles-ci entraînent des modifications qui engendrent des coûts supplémentaires. De plus, en matière de déploiement d'entreprise, les exigences de sécurité, d'efficacité de l'utilisation, de gouvernance et d'environnements multi-propriétaires sont souvent négligées par les développeurs dans leur course effrénée pour terminer au plus vite les projets. Cela a également un coût, d'autant plus élevé si l'environnement ne prend pas en charge la réutilisation.
- **Coûts liés aux changements de plate-forme et aux modifications.** Avec un secteur comme celui des Big Data, en perpétuelle mutation, il existe un écosystème de technologies de plate-forme de données qui évolue rapidement. Parmi les principaux risques et coûts que ces technologies entraînent pour les entreprises figure une approche de la gestion de Data Lake qui expose les entreprises au risque de devoir refactoriser et changer de plate-forme lorsque les technologies de plate-forme sous-jacentes évoluent. Il s'agit d'un problème courant lié au codage manuel, qui n'offre pratiquement aucune possibilité de réutilisation ou d'automatisation. C'est également vrai, mais dans une moindre mesure, pour les solutions de génération de code, qui prennent en charge un certain degré de réutilisation des données source (mais pas au niveau cible). Idéalement, vos éléments de plate-forme doivent être protégés et isolés des modifications apportées aux plates-formes de données sous-jacentes. Votre objectif est d'automatiser l'adaptation de vos logiciels à de tels changements, ce qui est impossible sans des capacités de réutilisation totale.
- **Coûts d'intégration.** Par intégration, nous désignons ici les coûts liés au fait de faire travailler ensemble différentes solutions de pointe. Si, par exemple, vous disposez d'un outil de préparation des données qui ne s'intègre pas avec votre catalogage des données, vous devrez supporter des coûts d'intégration et de maintenance pour ces deux

> **« J'ai besoin des fonctionnalités de réutilisation et de maintenance du code, avec la capacité d'obtenir des informations simples et rapides sur ce qui a été créé, et un déploiement rapide entre le développement et la production. »**
>
> **Ben,**
> Director of
> Platform Architecture

composants de votre solution. Il s'agit en réalité d'un autre exemple de coûts de développement, mais nous les avons séparés car ils concernent spécifiquement l'adoption de solutions ponctuelles, par opposition à des plates-formes.

- **Coûts d'administration, de mise à niveau et de maintenance.** Vous ne pouvez pas vous contenter de lancer une solution et de ne plus vous en occuper par la suite. Les autorisations des utilisateurs, les audits, la gouvernance, la surveillance de la conformité, les mises à niveau et les autres fonctions nécessitent une gestion continue, et ces heures de travail ont un coût. La mise à niveau de plusieurs solutions ponctuelles qui n'ont jamais été conçues pour fonctionner ensemble est un autre problème majeur en raison des interdépendances entre les différentes versions et configurations des technologies Big Data sous-jacentes. Les mises à niveau de plusieurs solutions ponctuelles peuvent s'avérer coûteuses et rapidement se traduire par des applications historiques trop complexes à mettre à niveau lorsque les fournisseurs ne prennent en charge aucune configuration de plate-forme compatible. Ces coûts doivent faire l'objet d'une estimation pour l'ensemble de votre solution.
- **Coûts de formation.** En cas d'utilisation de solutions ponctuelles, il est courant qu'un utilisateur soit contraint d'utiliser plusieurs outils pour réaliser un workflow. Cela nécessite que les utilisateurs se familiarisent avec plusieurs produits et/ou soient formés à les utiliser. Combien cela coûte-t-il ? Les éditeurs de logiciels facturent-ils des frais directs ? Combien d'heures de travail devrez-vous consacrer à la formation de votre personnel et/ou combien de temps faudra-t-il pour que vos utilisateurs utilisent efficacement ce logiciel ? Ces heures de travail représentent une perte de productivité et vous devez les convertir en coûts pour obtenir une image réaliste du coût total de possession. Gardez également à l'esprit que vous devrez parfois faire appel à des consultants externes pour mener à bien certains projets. Recherchez des outils qui vous aident à tirer parti des ressources existantes et des plates-formes qui partagent des métadonnées communes pour créer une expérience utilisateur plus unifiée et plus collaborative. Cela aidera votre organisation à optimiser l'efficacité du processus de bout en bout et à éliminer les coûteux retards dus à des problèmes de communication entre les analystes de données, les ingénieurs de données, les gestionnaires de données et les experts métiers.

Alors que les éléments précédents sont tous directement liés aux coûts, notamment au coût total de possession, nous devons aussi aborder les fonctionnalités qui optimisent la productivité. Ou, si l'on adopte le point de vue opposé, l'absence de ces fonctionnalités (ou leur insuffisance) présente des coûts par rapport à ce que vous pouvez raisonnablement attendre d'autres solutions. Ce domaine présente les considérations suivantes :

- **Coûts relatifs à la simplicité d'utilisation.** Sur le plan de la simplicité d'utilisation, le manque d'intégration des produits présente un coût. Bien qu'il s'agisse d'un domaine dans lequel les coûts sont difficiles à estimer, la simplicité d'utilisation a une incidence sur la productivité : la productivité des utilisateurs est proportionnelle à la convivialité de l'environnement.
- **Coûts relatifs à la collaboration.** Les solutions qui favorisent la collaboration et réduisent les délais de communication entre les équipes améliorent la productivité globale. Certains éléments de l'environnement de gestion de Data Lake, comme les outils de préparation des données, peuvent avoir des fonctionnalités collaboratives comme l'intégration des espaces de travail de projet, mais il existe d'autres fonctionnalités clés qui améliorent vraiment la collaboration. C'est notamment le cas des métadonnées communes entre les outils, qui ouvrent la voie à des interactions utilisateurs plus intelligentes comme la recherche, la découverte et les recommandations automatisées, et les comportements guidés intelligents tels que le croisement de deux ensembles de données disparates. La fourniture d'un glossaire métier, dont nous n'avons pas encore parlé, facilitera la collaboration avec les analystes et les experts métiers.

« **Nécessité de passer d'une approche qualitative fastidieuse à une approche systématique...** »

Ned,
SVP Enterprise
Data Management

Conclusion

Déterminer quelle approche choisir pour la gestion de Data Lake est une décision hautement stratégique qui a des répercussions sur le coût total de possession et le retour sur investissement. Il est important de sélectionner une approche qui optimise la productivité, l'efficacité et la maintenance, et qui pérennise votre investissement face aux évolutions des technologies Big Data sous-jacentes. Nous vous recommandons de commencer par envisager une solution hautement automatisée et intégrée qui réponde à l'intégralité de vos besoins, et de ne choisir un produit de pointe (cloisonné) pour la remplacer que si vous estimez que les capacités de la solution principale sont insuffisantes pour répondre à vos besoins.

POUR PLUS D'INFORMATIONS
Pour plus d'informations sur le sujet, visitez le site
www.bloorresearch.com/update/2313

À propos de l'auteur

PHILIP HOWARD
Research Director/Gestion de l'information

Philip Howard a fait ses débuts dans l'informatique en 1973 et a occupé des postes d'analyste système, de programmeur et de technico-commercial. Il a également exercé des responsabilités dans le domaine du marketing et de la gestion de produits pour différentes sociétés, notamment GEC Marconi, GPT, Philips Data Systems, Raytheon et NCR.

Après un quart de siècle passé à travailler pour un employeur, Philip Howard a créé sa propre société en 1992 et son premier client a été Bloor Research (à l'époque ButlerBloor). Il travaillait alors comme analyste associé pour cette dernière. Depuis cette période, il a continué d'entretenir des relations avec Bloor Research et il est à présent Research Director sur la gestion des informations.

La gestion des informations fait référence à la gestion, au déplacement, à la gouvernance et au stockage de données, ainsi qu'à l'accès et à l'analyse de ces données. Elle inclut diverses technologies, dont (entre autres) les bases de données et le data warehousing, l'intégration des données, la qualité des données, la gestion des données de référence, la gouvernance des données, la migration des données, la gestion des métadonnées, ainsi que la préparation et l'analyse des données.

Outre les nombreux rapports qu'il a rédigés pour le compte de Bloor Research, Philip Howard apporte également sa contribution régulière à ***IT-Director.com*** et ***IT-Analysis.com***. Il a été rédacteur en chef d'« ***Application Development News*** » et d'« ***Operating System News*** » pour le compte de Cambridge Market Intelligence (CMI). Il a également collaboré à de nombreux magazines et est l'auteur d'un certain nombre de rapports publiés par CMI et The Financial Times. Philip intervient régulièrement lors de conférences et autres événements dans toute l'Europe et l'Amérique du Nord.

Pendant son temps libre, Philip adore naviguer sur les canaux en péniche, skier, jouer au bridge (dont il est devenu un véritable expert) et aller au restaurant.

Présentation de Bloor

Bloor Research est l'une des principales sociétés de recherche, d'analyse et de conseil en Europe dans le domaine des technologies de l'information. En 2014, l'entreprise a célébré son 25e anniversaire. Nous expliquons comment accroître la flexibilité des systèmes informatiques d'entreprise au moyen d'une gouvernance, d'une gestion et d'une exploitation efficaces des données. Nous sommes réputés pour notre pertinence ; nos publications, de même que le contenu de nos communications, font preuve d'indépendance, d'intelligence et de cohérence, quel que soit l'aspect traité de l'industrie des technologies de l'information et des télécommunications (ICT). En formulant les choses de façon pertinente, nous pensons ainsi pouvoir :

- décrire la technologie au regard de sa valeur métier et des autres systèmes et processus avec lesquels elle interagit ;
- comprendre comment les nouvelles technologies et innovations s'intègrent aux investissements ICT existants ;
- offrir une vue d'ensemble du marché, expliquer toutes les solutions disponibles et la façon dont elles peuvent être évaluées plus efficacement ;
- filtrer les données superflues et faciliter les recherches d'informations ou d'actualités supplémentaires susceptibles de soutenir les investissements et l'implémentation ;
- veiller à ce que l'ensemble de notre contenu soit accessible par le biais du canal le plus approprié.

Notre société a été fondée en 1989. Depuis 25 ans, nous diffusons nos rapports à des utilisateurs et des fournisseurs informatiques du monde entier, par l'intermédiaire d'abonnements en ligne, de services d'études personnalisées, d'événements et de prestations de conseil. Nous mettons nos connaissances à votre disposition et espérons ainsi vous apporter une réelle valeur ajoutée.

Copyright et déni de responsabilité

Bloor Research International Ltd
20-22 Wenlock Road
LONDRES N1 7GU
Royaume-Uni

Tél : **+44 (0)20 7043 9750**
Internet : **www.Bloor.eu**
e-mail : **info@Bloor.eu**