L'EDW perdure

Le cœur battant du lac de données de

Avril 2017, dernière mise à jour mai 2018

Un livre blanc de

Dr Barry Devlin, 9sight Consulting
barry@9sight.com

Nous pouvons enfin dépasser le conflit entre entrepôt de données et lac de données ! Il ne s'agit plus de l'un contre l'autre mais plutôt de savoir comment ces deux concepts peuvent se compléter pour le bénéfice à la fois de l'entreprise et de l'informatique.

Tout d'abord, nous explorons comment prendre en charge de manière optimale différentes exigences commerciales et techniques en plaçant de manière appropriée des fonctionnalités telles que la préparation des données, l'archivage et l'accès commercial dans les deux environnements. Un modèle architectural simple définit ce que sont réellement les entrepôts et les lacs et comment ils se complètent. Cela démontre clairement la puissance d'une telle réflexion collaborative entre les approches traditionnelles et nouvelles.

Une brève description de la solution d'optimisation d'entrepôt de données d'entreprise de Hortonworks complète le document.

Parrainé par :

HORTONWORKS®

Contenu

Le comme deux combattants fatigués, l'entrepôt de données et le lac de données l'ont frappé sorti depuis une demi-décennie maintenant. Devons-nous remplacer notre entrepôt de données par un lac de données ? Un lac de données peut-il offrir une solution de business intelligence (BI) plus rentable qu'un entrepôt de données d'entreprise (EDW) ? Pourrions-nous transformer notre entrepôt de données propriétaire en un lac de données open source ? Faut-il le faire, et à quel prix ?

Ces questions et d'autres connexes passent à côté d'un point fondamental. Formulée en termes de l'un ou l'autre, l'implication est qu'un concept peut en déplacer un autre, qu'une approche peut être choisie à l'exclusion de toutes les autres. En fait, cette réflexion est erronée, motivée par des messages marketing obsolètes du début de cette décennie et, en fait, avant.

La réalité est que les entrepôts et les lacs de données sont des concepts largement complémentaires qui émergent de différents besoins commerciaux et possibilités technologiques. Vu de cette manière, deux possibilités surprenantes émergent. Premièrement, nous pouvons – et devrions – avoir les deux. Deuxièmement, la fonction peut être distribuée et redistribuée entre les deux environnements en fonction de la meilleure adéquation. Ensemble, ils promettent des améliorations de performances et des réductions de coûts : une BI meilleure, plus rapide, plus agile et plus rentable pour répondre aux besoins commerciaux en croissance rapide.

> Un entrepôt de données et un lac de données sont complémentaires dans une entreprise moderne.

Comment est-ce possible ? Le secret réside dans la compréhension des différences entre un entrepôt de données et un lac de données, d'abord en termes d'exigences commerciales et de possibilités techniques, puis à travers une image architecturale simple.

Ni les données de ton père, ni les idées de ta mère

Au bon vieux temps, toute l'aide à la décision et tous les rapports étaient basés sur des données provenant de je systèmes opérationnels que vous possédiez ou développiez. Ces données étaient gérées – depuis la préparation et le rapprochement, en passant par l'utilisation et la maintenance, jusqu'à l'archivage et la suppression – dans l'entrepôt de données où le service informatique se portait garant de leur qualité (relative). C'était peut-être cher, mais c'était faisable avec les volumes de l'époque, et de toute façon, il n'y avait pas beaucoup de choix. C'était peut-être un peu lent, mais c'était assez rapide pour la plupart des tâches professionnelles. pose.

Puis le monde a changé. Avec Internet et le commerce électronique, les affaires sont passées au temps réel. Les rapports de ventes d'une semaine ont été remplacés par des informations prédictives sur le comportement futur des clients. Ce tout nouveau système d'informations dépendait du comportement des clients sur le Web : likes, clics, commentaires, paniers abandonnés, relations, ventes incitatives et ventes croisées, etc. La BI est devenue analytique : l'accent est passé des rapports rétrospectifs et des états financiers précis aux évaluations probabilistes de qui pourrait faire quoi ensuite.

Les besoins des entreprises ont évolué et se sont élargis : plus rapides, plus larges, plus orientés vers l'avenir. Les mégadonnées, provenant d'abord des médias sociaux et des flux de clics, et plus récemment de l'Internet des objets, en sont devenues la base. Les désormais célèbres trois « V » – volume, vitesse et variété – ont bouleversé les équations de coûts d'un entrepôt de données traditionnel, conduisant à l'explosion de l'open source Hadoop. Notez cependant ici que la qualité et la fiabilité de ces nouvelles sources étaient et sont souvent médiocres. Et contrairement à l'opinion de certains créateurs de tendances, le besoin de données, de rapports et d'analyses BI à l'ancienne n'a pas disparu. Les données et les connaissances d'aujourd'hui doivent cohabiter avec celles de votre père et de votre mère.

C'est le défi de l'entreprise numérisée d'aujourd'hui. Nous devons disposer de toute urgence de nouvelles informations basées sur des sources de données modernes et principalement externes. Mais il faut aussi gérer l'entreprise dans le respect des impératifs juridiques et comptables, en s'appuyant sur les systèmes opérationnels et d'entrepôt de données développés au cours des trente dernières années. La reconfiguration ou le remplacement de ces systèmes existants serait une tâche complexe et coûteuse. En effet, pourquoi le feriez-vous ? Les entrepôts de données d'aujourd'hui sont plus puissants et sophistiqués que jamais. Des années d'investissement dans ces plates-formes et systèmes par les fournisseurs et le service informatique interne ont permis de produire des applications fonctionnelles et critiques pour l'entreprise.

Mais la nouvelle technologie peut-elle contribuer à améliorer ou à simplifier l'environnement existant ?

Du vieux vin dans des bouteilles neuves

Sur la base de nouvelles données et de nouvelles connaissances, un nouvel écosystème de gestion et de livraison de données basé sur Hadoop a émergé au cours de la dernière décennie : un lac de données (dont nous reviendrons sur la définition plus tard). Aujourd'hui, à mesure que cet écosystème mûrit, l'opportunité se présente de l'utiliser pour fournir des solutions meilleures et/ou moins coûteuses à certains des problèmes les plus insolubles de l'entreposage de données traditionnel :

> Un lac de données offre la possibilité de solutions meilleures et/ou moins coûteuses aux anciens problèmes d'entrepôt de données.

1. Préparation et enrichissement : la préparation des données pour l'entrepôt a longtemps été l'élément le plus complexe et le plus coûteux en termes de calcul d'un entrepôt de données. Traditionnellement appelé extraction, transformation et chargement (ETL), ce traitement est effectué soit dans un serveur ETL dédié, au sein de la base de données relationnelle de l'entrepôt (souvent appelée ELT - extraire, charger et transformer), ou dans une combinaison des deux. Dans de nombreux cas, ces systèmes sont basés sur des logiciels propriétaires, ce qui entraîne des coûts de licence élevés.
   De plus, lorsqu'il est effectué dans l'entrepôt, ce traitement est coûteux et peut interférer avec les tâches de BI ou d'analyse critiques pour l'entreprise.

   Pompage à travers le lac de données : la préparation et l'enrichissement des données dans l'environnement Hadoop mûrissent pour les sources de données externes, en commençant par les approches par lots et en évoluant vers des approches de streaming. Bien que certaines différences d'approche subsistent (les charges incrémentielles prédominent dans les entrepôts de données), la préparation des données sur Hadoop devient de plus en plus attrayante et puissante comme moyen de réduire le coût et l'impact des processus ETL effectués dans l'entrepôt de données lui-même.

2. Archivage : l'approche traditionnelle de l'archivage à partir d'entrepôts de données consiste à stocker sur bande magnétique. Tout en offrant le coût de stockage par téraoctet de loin le plus bas, les systèmes de bandes nécessitent souvent une intervention informatique manuelle ou au minimum des délais de montage physique des bandes pour la récupération, ce qui ralentit considérablement l'accès pour les utilisateurs professionnels. De plus, les utilisateurs doivent utiliser différents outils pour demander et/ou accéder aux données historiques, créant ainsi une barrière artificielle à leur utilisation quotidienne.

   Stockage dans le lac de données : l'environnement Hadoop est construit sur du matériel standard et offre ainsi un environnement d'archivage attractif. Bien que le coût par téraoctet soit nettement plus élevé que celui des bandes, le coût supplémentaire est plus que compensé par la facilité et la rapidité de récupération des données archivées directement par les utilisateurs professionnels sans aide. Grâce à la récupération dans le même langage (SQL) que l'utilisation en ligne, les utilisateurs professionnels perçoivent les données archivées comme aussi disponibles (peut-être avec un temps d'accès légèrement plus long) que les données en ligne, ce qui permet une meilleure utilisation des données de tendances historiques.

3. Accès : avec des quantités croissantes de données, principalement externes, ingérées dans l'environnement Hadoop, les utilisateurs professionnels sont confrontés à des difficultés pour accéder à ces données.
Jusqu'à récemment, une grande partie de cet accès se faisait via des outils qui dépassaient l'expérience des utilisateurs professionnels ou impliquaient des approches programmatiques plus adaptées aux développeurs informatiques. De plus, l'utilisation de données basées sur Hadoop avec des données traditionnellement trouvées dans les entrepôts ou les data marts pourrait impliquer de copier et coller des données d'un environnement à un autre, ce qui ajouterait des coûts et des efforts à la vie des utilisateurs professionnels.

Nager dans le lac de données : l'utilisation commerciale des données s'est centrée autour d'un paradigme « lignes et colonnes » depuis les premiers jours de la BI. Qu'il s'agisse de feuilles de calcul, de cubes multidimensionnels ou de requêtes SQL, offrir un tel accès aux données du lac de données est essentiel à leur utilisation généralisée par les utilisateurs professionnels « ordinaires ». Les données d'intérêt commercial étant désormais réparties sur deux environnements ou plus, il devient de plus en plus important de regrouper les données dans des emplacements physiquement distincts – une installation connue sous le nom de virtualisation des données.
De tels outils sont essentiels pour inciter les utilisateurs professionnels à se lancer dans le lac de données.

Ces exemples et d'autres émergents, tels que le streaming de données, le traitement hybride transactionnel/analytique (HTAP) et les séries temporelles, soulignent tous le fait qu'un lac de données a de la valeur à offrir à un entrepôt de données et vice versa. La complémentarité des deux environnements est évidente. Il est temps d'examiner de plus près la manière dont ils interagissent.

Un entrepôt au bord d'un lac

Source de tous les rapports BI, interrogation au début du millénaire. C'est original, création 1 au milieu des années 1980, l'entrepôt de données était devenu la référence depuis sa
Le principal objectif était d'offrir une base de données cohérente et réconciliée (souvent appelée version unique de la vérité) dans toutes les fonctions et tous les départements de l'entreprise. Bientôt, il a été déclaré que toutes les données d'aide à la décision devaient transiter par l'entrepôt pour le contrôle de la qualité et les analyses spécialisées (et aussi parce qu'il n'existait aucune autre plate-forme évidente pour un tel travail). Cette approche avait des impacts prévisibles sur les performances, l'agilité, les coûts, etc. Ces défis, ainsi que les nouveaux besoins commerciaux, ont conduit à des améliorations continues des logiciels sous-jacents, conduisant à l'environnement moderne et puissant que l'on connaît aujourd'hui.

Cependant, gardez à l'esprit le moteur initial mentionné ci-dessus : la réconciliation des données : elle est bien plus importante que l'idée que toutes les données doivent y résider.

Le concept de lac de données a été lancé pour la première fois en 2010 par James Dixon2.. Sa description initiale était simplement un vaste entrepôt de données brutes, motivé en partie par la croissance florissante du big data, mais aussi en réaction à la structuration souvent coûteuse et restrictive des données dans l'entrepôt.
et les datamarts. Bientôt, moi et d'autres3 avons critiqué le caractère vague de la définition, troublés par
la possibilité d'un « marécage de données » de données mal gérées. Dixon, entre autres, a élargi la portée du lac de données pour inclure toutes les données, y compris même celles traditionnellement stockées dans l'entrepôt de données, une approche qui suscite des inquiétudes supplémentaires quant aux coûts et aux défis liés à la « suppression et au remplacement » de l'écosystème de l'entrepôt de données. .

Malgré ces inquiétudes, la popularité du concept de lac de données a augmenté rapidement. Une enquête 9sight/EMA4 publiée en novembre 2016 montre que deux tiers des personnes interrogées ont déclaré avoir actuellement adopté une stratégie de lac de données, contre un peu plus de 50 % en un an et quart. En outre, près de 15 % des personnes interrogées lors de l'enquête de 2016 ont déclaré qu'un lac de données avait remplacé leur entrepôt de données. Cependant, le

L'enquête a montré une confusion considérable quant à ce que pourrait réellement être la définition d'un lac de données. Les composants du lac de données ayant obtenu des scores élevés comprenaient l'entrepôt de données, les données opérationnelles, les magasins de données départementaux et analytiques, ainsi que les bacs à sable d'exploration de données.

Alors, qu'est-ce qu'un lac de données ? Et pendant que nous sommes en mode définition, qu'est-ce qu'un entrepôt de données ?

Définition de la terminologie : entrepôt de données et lac de données

Après plus de trente ans, la définition conceptuelle d'un entrepôt de données est stable, même si en termes logiques ou fonctionnels, certaines différences sont évidentes* . Un aperçu de haut niveau est présenté dans l'encadré d'accompagnement, basé sur mon livre de 2013 « Business unIntelligence »5 . La définition reflète l'évolution du concept d'entrepôt de données au cours de ses premières années, avec des composants particuliers optimisés pour des objectifs spécifiques basés sur l'évolution des caractéristiques des bases de données relationnelles sur trois décennies. L'EDW, avec son rôle de nettoyage et de rapprochement des données provenant de nombreuses sources, est essentiel pour comprendre la différence entre un entrepôt de données et un lac de données.

> Entrepôt de données : un environnement de collecte, de gestion et de stockage de données pour l'aide à la décision, composé de :
>
> Entrepôt de données d'entreprise (EDW) : un magasin détaillé, nettoyé, réconcilié et modélisé de données historiques interfonctionnelles
>
> Datamarts : sous-ensembles d'aide à la décision données optimisées et stockées physiquement pour des utilisations spécifiques par les hommes d'affaires

L'objectif principal d'un entrepôt de données est donc de fournir un ensemble de données fiables et cohérentes aux utilisateurs professionnels pour les aider à prendre des décisions, en particulier pour les actions juridiquement pertinentes, le suivi des performances et l'identification des problèmes. Ces données détaillées proviennent des systèmes opérationnels, mais peuvent être subdivisées ou résumées selon les besoins au moment où un utilisateur professionnel les consulte.

En revanche, un lac de données est souvent défini en termes d'attributs qui le caractérisent, comme le montre l'extrait suivant, légèrement édité à partir du billet de blog de Shaun Connolly de 20146 :

« Un Data Lake se caractérise par trois attributs clés :

1. Collectez tout : contient toutes les données, à la fois les sources brutes et les données étendues. périodes de temps et toutes données traitées
2. Plongez n'importe où : permet aux utilisateurs de toutes les unités commerciales d'affiner, par exemple explorer et enrichir les données selon leurs conditions
3. Accès flexible : permet plusieurs modèles d'accès aux données sur une infrastructure partagée : par lots, interactif, en ligne, recherche, en mémoire, etc.

> Lac de données : un magasin de données conçu pour l'ingestion et le traitement de toutes données brutes provenant de plusieurs sources sans structuration préalable selon un modèle préféré. Dans ce magasin, les données peuvent être consultées, formatées, traitées et gérées selon les besoins à des fins commerciales ou techniques.

Le défi de cette définition est qu'elle implique que le lac de données contient tous les éléments de données imaginables, permet le traitement si nécessaire et peut essentiellement répondre à tous les besoins commerciaux ou techniques. Je propose la définition la plus limitée et la plus utile à droite, basée sur les besoins originaux notés par James Dixon et axée sur les fonctionnalités en dehors du cadre d'un entrepôt de données. Bien que d'autres experts puissent être en désaccord avec cette dernière restriction, un avantage évident est qu'elle concentre les efforts dans les domaines qui profitent le plus à la majorité des entreprises qui ont déjà investi dans l'entreposage de données pour répondre à leurs besoins existants.

---

* En particulier, l'entrepôt de données de Kimball utilise un modèle de données dimensionnel (schéma en étoile) comme base. tion pour une analyse « tranche et dés ». Une telle construction apparaît comme un datamart dans la définition ci-dessus.

Cette division du travail permet la création d'une architecture logique simple, comme le montre la figure 1, qui positionne l'entrepôt de données et le lac de données les uns par rapport aux autres, définissant des rôles qui peuvent être compris par l'entreprise ainsi que par l'informatique.

Le bloc labellisé fonctionnel est au cœur de la gestion et de la gestion d'une entreprise selon les pratiques éthiques, juridiques et comptables. Cela commence par la collecte ou la création de transactions juridiquement contraignantes qui représentent de véritables activités commerciales, comme la création d'un compte client ou l'acceptation d'une commande. Il passe par les processus opérationnels qui génèrent de la valeur et se termine par les processus informationnels utilisés pour suivre les progrès et résoudre les problèmes. Ainsi, cela s'étend de la programmation Cobol dans les années 1950 aux entrepôts de données « typiques » et aux outils de BI d'aujourd'hui. L'exactitude et la cohérence des données utilisées sont essentielles à l'informatique fonctionnelle : si les données sont erronées, l'entreprise s'effondre ou le régulateur intervient.
Avant l'ère d'Internet, ces transactions étaient uniquement utilisées par les entreprises et gérées par l'informatique.

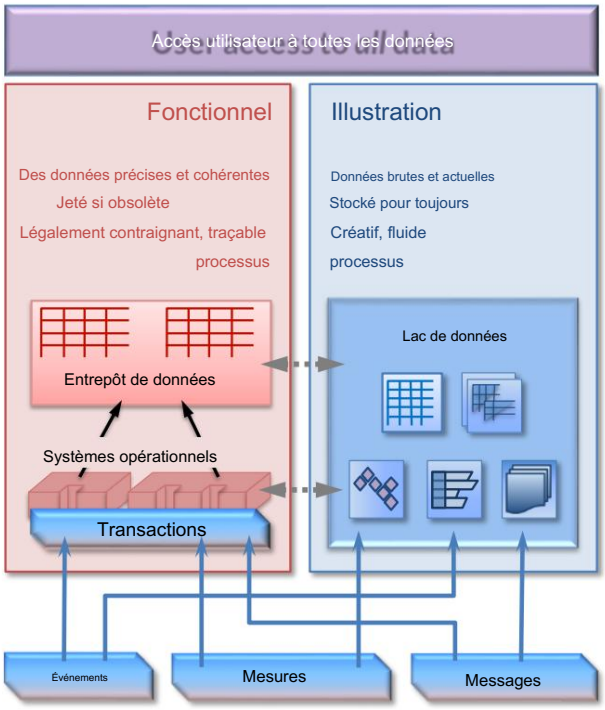

Figure 1 : L'entrepôt au bord du lac

Le commerce électronique, les médias sociaux et l'IoT montrent qu'il existe des données « plus brutes » à partir desquelles les transactions découlent. Ces données/informations, désormais toutes numérisées et potentiellement collectées, sont constituées d'événements (par exemple un clic sur un site Web), de mesures (la vitesse de votre voiture) et de messages (tout, des Tweets aux vidéos). Ces données soutiennent des processus illustratifs qui permettent de déduire ce qui se passe dans le « monde réel » et constituent la base d'analyses prédictives et prescriptives. L'actualité et le caractère brut des données sont essentiels à l'informatique illustrative ; les retards ou les résumés dégradent souvent la valeur analytique.

Ces objectifs fonctionnels et illustratifs, avec leurs caractéristiques et utilisations de données opposées, conduisent à une architecture qui définit les rives du lac de données. Données brutes : dans le sous forme d'événements, de mesures et de messages – est ingéré dans les systèmes informatiques de l'entreprise. De grandes quantités de données brutes peuvent être stockées dans le lac de données comme base d'analyse. Les systèmes opérationnels traditionnels transforment les données brutes en transactions juridiquement contraignantes de l'entreprise et les rendent disponibles pour la prise de décision via l'entrepôt de données. Cette séparation des préoccupations permet de séparer les données et les processus qui doivent être bien gérés pour la continuité des activités et la légalité de ceux qui nécessitent moins de gestion mais permettent plus de créativité. Un lac de données répond à ces derniers besoins, un entrepôt au premier. Pour les utilisateurs professionnels, cette séparation du stockage est masquée et gérée par des outils de virtualisation des données et des approches basées sur les métadonnées. Des liens profonds (les flèches en pointillés) existent entre les deux environnements pour des besoins commerciaux spécifiques tels que les approches analytiques prescriptives qui sont de plus en plus répandues.

Notez que cette image architecturale n'implique aucun placement physique de l'un ou l'autre boîtier sur site, dans le cloud (privé ou public) ou toute combinaison de ceux-ci. En fait, dans le
Dans un environnement cloud émergent, le placement le plus probable est une approche hybride sur site et cloud en fonction des sources des principaux types de données impliquées.

Avec l'architecture équilibrée entre entrepôt et lac de données présentée ici, les solutions de traitement des données, d'archivage et d'accès décrites dans la section précédente, ainsi que d'autres possibilités d'utiliser le lac de données pour améliorer, prendre en charge ou étendre l'entrepôt de données, peuvent être efficaces. -scientement livré sur le matériel et les logiciels les moins coûteux du lac de données.

Solution d'optimisation Hortonworks EDW

En février 2017, Hortonworks a déployé la première phase d'une solution d'optimisation Enterprise Data Warehouse†
je (EDW) pour faciliter l'utilisation de la fonction de lac de données pour la prise en charge de la création, de la gestion et de l'utilisation des données EDW. Cette approche représente l'aboutissement d'une évolution continue de la réflexion sur les lacs de données au sein de la communauté Hadoop.
Il s'agit d'une évolution très bienvenue, qui fait passer le débat de l'opposition entre lac et entrepôt à une position plus réaliste consistant à utiliser les atouts de l'écosystème Hadoop pour répondre aux nouveaux besoins commerciaux, ainsi qu'à tirer parti et à améliorer les investissements technologiques existants.

Conformément à sa stratégie de longue date consistant à fournir des distributions entièrement intégrées et testées d'un ensemble de composants de l'écosystème Hadoop, Hortonworks rassemble désormais des composants supplémentaires provenant de fournisseurs partenaires pour combler le fossé entre l'entrepôt de données et le lac de données et fournir les fonctionnalités dans les trois domaines décrits aux pages 3 et 4 : préparation et enrichissement, archivage et accès.

Préparation et enrichissement des données de l'entrepôt

La solution d'optimisation EDW propose des flux de travail ETL par lots et en streaming simples par glisser-déposer en incorporant DMX-h de Syncsort, qui offre un accès aux données provenant de sources multiples, notamment des bases de données relationnelles, des magasins NoSQL, des fichiers et bases de données mainframe, etc. ., et génère une fonction ETL hautement évolutive dans Hive et Horton-

fonctionne Data Platform (HDP).

La plateforme Hadoop offre une grande flexibilité quant au type de données pouvant être stockées
là. L'EDW traditionnel se concentre étroitement sur les données structurées, avec une conception initiale stricte, connue sous le nom de « schéma en écriture ».
Hadoop, quant à lui, peut stocker n'importe quelle forme de données, structurées, semi-structurées, non structurées et s'associera à une structure lors de l'accès – « schéma en lecture ». Cela permet
les sources de données modernes qui ne s'intègrent pas facilement dans l'EDW, comme les flux de clics, les journaux Web, les données des appareils, etc., doivent être plus facilement préparées et enrichies ici, ainsi que le stockage et la prise en charge des données précédemment archivées.

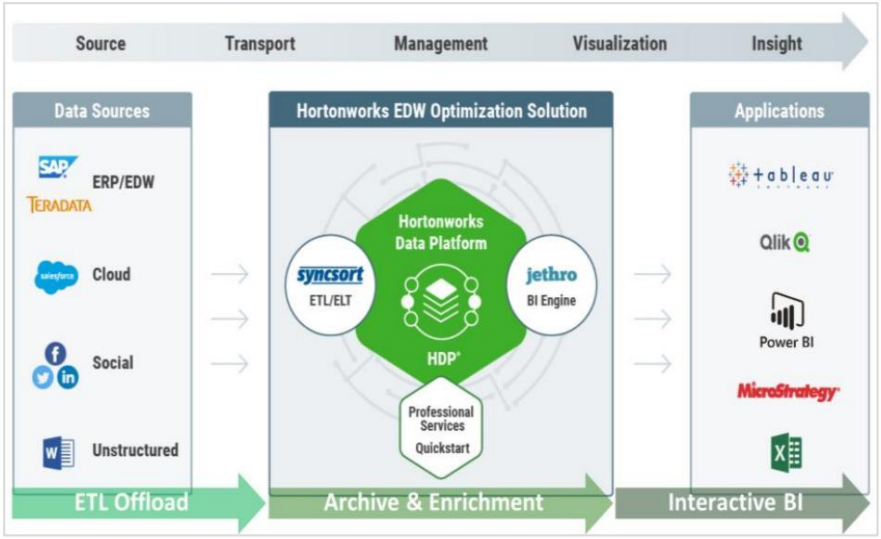

Figure 2 : EDW de Hortonworks
Solution d'optimisation

Dans les cas où l'ETL existant est effectué dans l'entrepôt de données, cette approche peut déplacer plus de 50 % du traitement hors de la plateforme de l'entrepôt de données, ce qui entraîne des améliorations significatives des performances et des accords de niveau de service (SLA). Lorsque l'ETL existant est réalisé sur des systèmes propriétaires, ces outils existants peuvent être progressivement supprimés au fil du temps pour générer des économies de coûts intéressantes.

† Hortonworks utilise le terme « Entrepôt de données d'entreprise » pour inclure à la fois les fonctions de rapprochement et de magasin de données décrites à la page 5 afin de souligner la nature globale de l'entrepôt de données à l'échelle de l'entreprise.

Archivage des données de l'entrepôt

Stocker des données rarement utilisées (ou froides) dans un entrepôt de données hautes performances constitue une mauvaise utilisation d'une ressource coûteuse. L'archivage traditionnel déplace les données anciennes et rarement utilisées sur bande. Avec un lac de données, ces données peuvent être déplacées vers Hadoop. En outre, la même approche peut être utilisée pour toutes les données rarement utilisées dans l'entrepôt, y compris les données brutes provenant de l'Internet des objets, des flux de clics, des médias sociaux et d'autres sources.

Pour Syncsort DMX-h, l'entrepôt de données n'est qu'une autre source, donc les avantages répertoriés dans la section précédente s'appliquent également ici. De même, toutes les données archivées bénéficient des composants d'accès répertoriés ci-dessous, ce qui rend ces données facilement disponibles à tout moment pour analyse :
pour l'utilisateur, il semble qu'il n'ait jamais quitté l'entrepôt de données.

Accéder vers Données du lac de données

Hive 2.0 avec LLAP (Live Long and Process) offre des analyses SQL évolutives et une mise en cache intelligente en mémoire en moins d'une seconde. Le résultat est une amélioration des performances 26 fois supérieure à celle de Hive 1.0, permettant de véritables requêtes interactives sur les données stockées dans HDFS. En conséquence, les outils traditionnels d'accès aux données et de requête du monde des entrepôts de données, tels que Qlik, Tableau, etc., peuvent utiliser directement les données du lac de données comme base pour les requêtes et les analyses.

En collaboration avec Jethro Data, la solution d'optimisation EDW accélère encore l'accès aux données grâce à une indexation intelligente. Cette nouvelle fonctionnalité indexe automatiquement chaque colonne, regroupe les données pour les cubes OLAP et met en cache les données hautement consultées, le tout grâce à une conduite autonome qui ne nécessite aucune ingénierie des données de la part du service informatique ou des utilisateurs. L'approche innovante de Jethro optimise la manière dont les données sont accessibles, offrant des performances à grande échelle pour des milliers d'utilisateurs simultanés avec un temps de réponse en secondes et une évolutivité permettant de traiter des milliards de lignes pour les requêtes.

Conclusions

Malgré trois décennies d'histoire, l'entrepôt de données reste un élément central.
Dnent dans toute architecture d'aide à la décision. Au cours des cinq dernières années, un nouveau
Un composant – le lac de données – a été introduit dans le mix. Considéré au départ comme très compétitif par rapport à l'entrepôt de données, une réflexion plus évoluée le place comme un partenaire égal.
L'entrepôt de données conserve la responsabilité des données réconciliées et juridiquement fondamentales nécessaires au fonctionnement et à la gestion responsable de l'entreprise. Le lac de données, quant à lui, offre un endroit pour stocker les données brutes et les traiter de manière innovante et en constante évolution.

De plus, le lac de données offre un environnement pour décharger les données stocker une fonction qui posait problème dans le passé. De telles fonctions, telles que la préparation et l'archivage des données, peuvent avoir un impact sur les performances et les SLA sur l'entrepôt de données et peuvent être exécutées à moindre coût dans le lac de données. En déplaçant cette fonction hors de l'entrepôt de données, la durée de vie de l'environnement existant peut être prolongée ou les coûts d'exploitation réduits.

> Les performances et l'utilisation d'un entrepôt de données peuvent être optimisées en déplaçant certaines fonctionnalités vers une autre plateforme appropriée telle que le lac de do

La solution d'optimisation Hortonworks EDW est un ensemble intégré de composants de l'écosystème Hadoop et de fournisseurs de logiciels partenaires qui répondent à trois divisions :

aspects subtils mais interdépendants de l'utilisation du lac de données pour améliorer et étendre l'entrepôt de données. Premièrement, il prend en charge le déchargement de la préparation des données (ETL) de l'entrepôt de données ou des outils existants afin de réduire les coûts et d'améliorer les performances. Deuxièmement, il permet d'archiver les données de l'entrepôt sur une boutique en ligne plutôt que sur bande, permettant aux utilisateurs d'accéder plus rapidement et plus simplement à ces données historiques. Troisièmement, il offre aux utilisateurs professionnels la possibilité d'utiliser des outils BI familiers pour accéder et utiliser toutes les données du lac de données, y compris les données archivées. Nous pouvons envisager que d'autres fonctions soient ajoutées à l'avenir.

Cette évolution de l'architecture d'un entrepôt à un lac puis à un entrepôt au bord d'un lac promet de fournir aux utilisateurs professionnels une fonction d'illustration multi-environnements indispensable pour explorer les données de manière créative, ainsi que d'optimiser l'environnement de l'entrepôt pour se concentrer sur les besoins fonctionnels de fournir des données correctes et cohérentes. des données cohérentes pour se conformer aux besoins commerciaux, juridiques et réglementaires. De plus, cette intégration et cette connexion des lacs et des entrepôts offrent la possibilité de faire encore plus avec plus de données, créant ainsi de nouvelles opportunités basées sur les données pour les entreprises traditionnelles et basées sur Internet.

Le Dr Barry Devlin est l'une des plus grandes autorités en matière de business insight et l'un des fondateurs de l'entreposage de données, ayant publié le premier article architectural sur le sujet en 1988. Avec plus de 30 ans d'expérience en informatique, dont 20 ans chez IBM en tant qu'ingénieur émérite, il est un homme très respecté
analyste, consultant, conférencier et auteur du livre phare « Data Ware-house—from Architecture to Implementation » et de nombreux livres blancs.
Son nouveau livre, « Business unIntelligence—Insight and Innovation Beyond An-alytics and Big Data » (http://bit.ly/BunI-Technics) a été publié en 2013.

Barry est fondateur et directeur de 9sight Consulting. Il se spécialise dans les implications humaines, organisationnelles et informatiques des solutions de business insight approfondies qui combinent des environnements opérationnels, informationnels et collaboratifs. Tweeter régulier, @BarryDevlin, et contributeur à de nombreuses publications, Barry est basé à Cape Town, en Afrique du Sud et opère dans le monde entier.

[1] Devlin, BA et Murphy, PT, « Une architecture pour un système d'entreprise et d'information », IBM Systems Journal, volume 27, n° 1, page 60 (1988) http://bit.ly/EBIS1988

[2] Dixon, J. « Blog de James Dixon : Pentaho, Hadoop et Data Lakes », (octobre 2010), https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/

[3] Stonebraker, M. « Pourquoi le « lac de données » est vraiment un « marais de données » », (décembre 2014), http://cacm.acm.org/blogs/blog-cacm/181547-pourquoi-le-lac-de-données-est-vraiment-un-marais-de-données/fulltext

[4] Myers, J., Wise, L. et Devlin, B., « Charting the Expanding Horizons of Big Data », (novembre 2016), http://bit.ly/BD-survey16

[5] Devlin, B., « Business unIntelligence », (2013), Technics Publications LLC, http://bit.ly/BunI_Book

[6] Connolly, S., « Enterprise Hadoop et le parcours vers un lac de données » (mars 2014), https://hortonworks.com/blog/enterprise-hadoop-lac-de-données-voyage/