Émergence de nouvelles bonnes pratiques pour le Big Data

Un livre blanc de Kimball Group
rédigé par Ralph Kimball

KIMBALL GROUP
Consulting | Kimball University

La révolution du big data est déjà en marche. Nous savons que ce n'est pas le volume des données en jeu qui fait l'intérêt du big data. D'ailleurs, le big data s'écarte sensiblement des données textuelles et numériques familières stockées dans des bases de données relationnelles et analysées avec SQL depuis plus de 20 ans. Le format et le contenu du big data vont de formats structurés relationnels à du texte libre non structuré aux vecteurs, matrices, images et collections de paires nom-valeur.

Le système a subi un premier grand choc lorsqu'il est devenu évident que les bases de données relationnelles standards et SQL n'étaient pas en mesure de stocker ni de traiter le big data, et qu'elles atteignaient leurs limites en termes de capacité et d'évolutivité. Non seulement les formats de données vont au-delà de la portée des bases de données relationnelles, mais une grande partie du traitement nécessite une logique itérative, des ramifications complexes et des algorithmes d'analyse spécifiques. SQL est un langage déclaratif doté d'une syntaxe puissante, mais figée.
Le big data requiert généralement des langages procéduraux et la possibilité de programmer avec une nouvelle logique floue.

Le deuxième grand choc subi par le système correspond au délaissement du reporting basé sur la permutation des axes d'analyse reposant sur des filtres simples et l'agrégation en faveur de l'analyse. Les rapports, tableaux de bord et requêtes ad hoc auront toujours de l'importance, mais on exploite mieux le big data en « ratissant » de vastes jeux de données non filtrés combinant les données historiques et en temps réel.

Pour finir, le troisième choc a été la prise de conscience que la valeur du big data augmente à vive allure à mesure que la latence diminue et que les données sont livrées plus rapidement. Quelle soit multipliées par 10 ou 100, l'amélioration des performances génère des opportunités d'analyse différentes sur le plan qualitatif qui se traduisent généralement par l'augmentation du chiffre d'affaires et des bénéfices.

Tous ces éléments ont contribué au dynamisme du marché porté par la technologie, avec deux tendances de développement principales : les bases de données relationnelles étendues et Hadoop. J'ai décrit dans le détail ces architectures dans un précédent livre blanc parrainé par Informatica. Ces deux architectures s'efforcent de traiter les enjeux du big data dont il est question ci-dessus.

Le marché du big data est loin d'avoir atteint sa maturité, mais nous disposons à présent d'une vision sur plusieurs années et des meilleures pratiques propres au big data. Ce livre blanc aborde ces meilleures pratiques en suivant une ligne médiane entre les observations d'ordre général et les aspects techniques approfondis spécifiques à un outil unique.

Il est important de savoir que nous disposons d'un ensemble de bonnes pratiques bien rodé, développé pour les data warehouses d'entreprise basés sur des relations que le big data doit exploiter. En voici la liste :

- Orienter le choix des sources de données qui alimentent le data warehouse d'entreprise selon les besoins métiers
- Ne jamais perdre de vue la simplicité et les performances de l'interface utilisateur

La liste suivante répertorie les meilleures pratiques de data warehouse d'entreprise particulièrement pertinentes pour le big data :

- Penser « dimensions » : diviser le monde en dimensions et en faits
- Intégrer les sources de données distinctes aux dimensions dites conformes (partagées)
- Établir un suivi des écarts de temps dans les dimensions qui évoluent lentement
- Ancrer toutes les dimensions au moyen de clés de substitution durables

La suite de ce livre blanc répartit les meilleures pratiques du big data en quatre catégories : gestion des données, architecture des données, modélisation des données et gouvernance des données.

Meilleures pratiques de gestion appliquées au Big Data

Les meilleures pratiques suivantes s'appliquent à la gestion globale d'un environnement Big Data.

Structurez vos environnements Big Data autour des analyses et non des requêtes ad hoc ou des rapports standards. Chaque étape du parcours des données depuis le contenu source jusqu'à l'écran de l'analyste doit prendre en charge des opérations d'analyse complexes implémentées en tant que fonctions définies par l'utilisateur ou via un environnement de développement reposant sur des métadonnées programmable pour chaque type d'analyse. Cette approche inclut les outils de chargements nettoyages, intégratations, interfaces utilisateurs et outils de Business Intelligence (BI). Cette meilleure pratique ne recommande pas de rejeter votre environnement existant, mais plutôt de l'étendre afin qu'il prenne en charge les nouvelles exigences d'analyse. Consultez la section ci-après relative aux meilleures pratiques appliquées à l'architecture.

À ce stade, n'essayez pas de créer un environnement Big Data figé. L'environnement de big data évolue à présent trop rapidement pour permettre la mise en place de fondations durables. Envisagez plutôt d'intégrer les facteurs perturbateurs arrivant de toutes parts : nouveaux types de données, concurrence, programmation, matériels, technologies réseau et services proposés par des centaines de nouveaux fournisseurs en big data. Dans un avenir proche, veillez à maintenir l'équilibre entre les diverses approches de mise en œuvre comme Hadoop, un grid traditionnel, l'optimisation Pushdown dans un SGBDR, l'informatique sur site, le cloud computing et même votre infrastructure mainframe.

Aucune de ces approches ne sera le grand vainqueur à long terme. Les fournisseurs de solutions PaaS (Platform as a service, plate-forme en tant que service) proposent une option attrayante capable de réunir un ensemble d'outils compatibles. D'une manière similaire, la majeure partie de l'architecture et de la programmation du système peut être définie dans une couche au-dessus des choix de déploiement spécifiques, ce qui constitue un réel avantage des environnements de développement reposant sur des métadonnées.

> Par exemple, utilisez HCatalog dans l'environnement Hadoop pour fournir une couche d'abstraction située au-dessus de l'emplacement du stockage et du format des données. Cette démarche permet de ne pas modifier les scripts Pig, par exemple, lorsque les emplacements et les formats subissent des modifications.
>
> Autre exemple : considérez Hadoop comme un environnement générique flexible adapté à de nombreuses formes de traitement ETL où l'objectif est d'apporter suffisamment de structure et de contexte au big data pour qu'il puisse être chargé dans un système de gestion de base de données relationnelle (SGBDR). Dans Hadoop, ces mêmes données peuvent être accessibles et transformées au moyen de code Hive, Pig, HBase et MapReduce rédigé dans plusieurs langages, même simultanément.

Cette démarche exige avant tout de la flexibilité. Partez du principe que vous allez reprogrammer et réhéberger vos applications big data dans les deux prochaines années. Choisissez des approches qui autorisent la reprogrammation et le réhébergement. Réfléchissez à l'utilisation d'un environnement de développement sans code basé sur des métadonnées pour augmenter la productivité et protéger votre entreprise des changements technologiques sous-jacents.

Acceptez les bacs à sable en silos et prenez l'habitude d'utiliser les résultats obtenus dans le cadre de la phase de production. Autorisez les experts en données à mener leurs expérimentations et créer leurs prototypes à l'aide de leurs langages et environnements de programmation préférés. Ensuite, après validation de principe, reprogrammez et/ou reconfigurez systématiquement ces implémentations avec le renfort d'une équipe informatique.

> Par exemple, vous disposez d'un environnement de production MatLab avec PostgreSQL ou SAS dans un SGBDR Teradata pour la programmation d'analyses spécifiques, mais vos experts en données créent leurs validations de principe dans divers langages et architectures. Le point clé sous-entendu, c'est que l'équipe informatique doit tolérer sans distinction de caractéristiques les technologies utilisées par les experts et se préparer, dans bien des cas, à procéder à une nouvelle implémentation de leur travail dans un ensemble de technologies standards qui seront prises en charge sur le long terme.

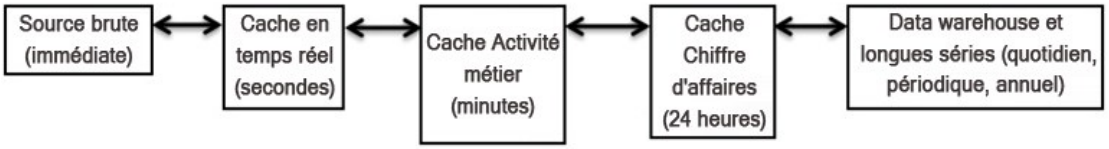
Par exemple, votre environnement de développement en test peut être une combinaison de transformations ETL associée à un code R personnalisé accédant directement à Hadoop, mais sous le contrôlé d'Informatica PowerCenter. Ensuite, lorsque l'expert en données est prêt à remettre la validation de principe, une grande partie de la logique pourrait être immédiatement redéployée sous PowerCenter afin de s'exécuter dans un environnement grid évolutif, hautement disponible et sécurisé.

Commencez progressivement avec une application big data simple dédiée à la sauvegarde et à l'archivage. Lorsque vous commencez à utiliser votre programme et lorsque vous recherchez des exemples d'utilisation intéressants, sans risque ou presque, puis lorsque vous réunissez les compétences requises en big data, pensez à exploiter Hadoop comme technologie de sauvegarde et d'archivage flexible à faible coût. Hadoop peut stocker les données et les extraire quel que soit leur format : du format non structuré au format spécifique très structuré. Cette approche peut également vous permettre de faire face au « déclin » des applications d'origine, qui risquent de ne pas être disponibles à l'avenir (peut-être en raison de restrictions de licence), pour que vous puissiez récupérer des données issues de ces applications dans votre propre format documenté. Pour finir, n'oubliez pas que même Hadoop consomme des ressources et de l'argent. En conséquence, chaque fois que des données sont stockées dans Hadoop, leur conservation doit être étudiée à l'avance afin de pouvoir purger ou archiver les dossiers et jeux de données HDFS à l'extérieur d'HDFS dans le but de réduire les coûts de stockage lorsque la période de conservation arrive à son terme.

Meilleures pratiques d'architecture appliquées au Big Data

Les meilleures pratiques suivantes concernent la structure et l'organisation en général de votre environnement Big Data.

Prévoyez une « autoroute de l'information » logique dotée de multiples caches pour gérer les temps de latence. N'implémentez physiquement que les caches adaptés à votre environnement. Cette autoroute de l'information peut disposer d'un maximum de cinq caches de gestion de la latence de données, chacun associé à des avantages et inconvénients en matière d'analyse :



* Applications sources brutes : détection des fraudes à la carte bancaire, traitement immédiat des événements complexes, notamment la stabilité du réseau et la détection de cyberattaques.

* Applications en temps réel : sélection de publicités sur pages Web, promotions tarifaires personnalisées, contrôle des jeux en ligne, surveillance prédictive et proactive sous diverses formes.

- Applications d'activités métiers : tableaux de bord des indicateurs de performance clé à faible latence pour les utilisateurs, suivi des dossiers d'incident, reporting d'alerte du traitement des événements complexes, portails et tableaux de bord de service client, applications de ventes mobiles.

* Applications décisionnelles : reporting tactique, suivi des promotions, corrections au stade intermédiaire basées sur le buzz des réseaux sociaux. Tableaux de bord pour la pratique courante des cadres dirigeants qui consultent la synthèse des événements survenus dans l'entreprise au cours des dernières 24 heures.

- Applications de data warehouse d'entreprise et prolongées dans le temps : toutes les formes de reporting, les requêtes ad hoc, l'analyse historique, la gestion des données de référence, la dynamique temporelle à grande échelle, l'analyse de la chaîne Markov.

Chaque cache qui existe dans un environnement donné est un cache physique distinct des autres. Les données se déplacent en provenance de la source brute tout au long de cette autoroute par le biais des processus ETL. Il est possible qu'il existe plusieurs chemins entre la source brute et les caches intermédiaires. Par exemple, les données peuvent accéder au cache en temps réel pour offrir une interface utilisateur sans latence, et dans le même temps être extraites directement dans un cache de synthèse journalier, similaire à une base tampon (ODS) classique. Ensuite, les données de cette base tampon pourraient alimenter le data warehouse d'entreprise. Les données circulent également dans le sens inverse le long de cette autoroute de l'information. Consultez la section « Implémentation des retours de données » plus loin dans ce document.

La majeure partie des données qui emprunte cette autoroute doit utiliser un format non relationnel allant du texte non structuré aux données complexes multi-structurées, telles que les images, tableaux, graphiques, liens, matrices et paires nom-valeur.

Utilisez les analyses big data en tant que « extracteurs de faits » pour déplacer les données vers le cache suivant. Par exemple, l'analyse de tweets non structurés peut générer un ensemble de mesures numériques des impressions, comme la part de voix, l'engagement des personnes, la portée des conversations, les soutiens actifs, les influenceurs et leur impact, le taux et l'heure de résolution, le niveau de satisfaction, les tendances actuelles, la teneur des impressions qui se dégagent et l'impact des idées. Renseignez-vous également sur la technologie Splunk, qui extrait des fonctions et indexe de nombreuses formes de données système non structurées ; la technologie Kapow, qui permet d'extraire diverses formes de données Web des blogs, forums de discussion, sites Web et portails ; et naturellement Informatica HParser, capable d'extraire des faits et des dimensions à partir de documents texte non structurés, de documents XML et journaux Web multi-structurés, de structures standards comme les données de marché SWIFT, FIX, CDR, HL7, HIPAA, etc.

Utilisez l'intégration du big data pour créer des écosystèmes complets qui intègrent des données SGBDR conventionnelles structurées, les documents papier, les e-mails et les données des réseaux sociaux internes orientés métier. Le big data présente l'avantage de pouvoir intégrer des sources de données disparates aux modalités très différentes. Nous recevons des flux de données de nouveaux canaux générateurs de données comme les réseaux sociaux, les appareils mobiles et les processus d'alerte automatisés. Imaginez une grande institution financière devant gérer des millions de comptes, les dizaines de millions de documents papier associés ainsi que des milliers de professionnels internes et externes, qu'ils soient partenaires ou clients. Configurez à présent un « réseau social » sécurisé de toutes les parties en présence afin de communiquer les décisions à mesure qu'elles sont prises. La majeure partie de cette communication est importante et doit être enregistrée dans un format qui prend en charge les requêtes. Capturez toutes les informations dans Hadoop, dimensionnez-les en vous aidant des meilleures pratiques de modélisation ci-dessous, utilisez-les dans le cadre de vos activités, puis sauvegardez-les et archivez-les.

Agissez en faveur de la qualité des données tout au long de l'autoroute de l'information. Il s'agit ici du compromis classique entre latence et qualité. Les analystes et utilisateurs métiers doivent accepter la réalité, à savoir que les données de très faible latence (immédiates) sont inévitablement de mauvaise qualité car il existe des limites quant aux opérations de nettoyage et de diagnostic réalisables dans de brefs intervalles de temps. Il est néanmoins possible d'exécuter des tests et d'apporter des corrections au contenu des champs individuels tout en exploitant les débits de transfert les plus élevés. Les tests et corrections effectués sur les relations structurelles entre les champs et les sources de données sont invariablement plus lents. Les tests et corrections impliquant des règles métiers complexes peuvent être instantanés (comme le classement de dates dans un certain ordre) ou être extrêmement longs sans raison (comme les temps d'attente pour savoir si un seuil d'événements inhabituels a été dépassé). Pour finir, les processus ETL plus lents, comme ceux qui alimentent le cache de synthèse journalier, reposent fréquemment sur des données plus complètes dans lesquelles sont éliminés, par exemple, les jeux de transaction incomplets et les transactions rejetées. Dans ce type de cas, les flux de données instantanés ne possèdent tout simplement pas l'information souhaitée.

Exécutez des opérations de filtrage, nettoyage, élimination, partage rapprochement, association et diagnostic dès que possible. Ces actions sont le corollaire de la meilleure pratique précédente. Chaque étape de l'autoroute de l'information offre plus de temps pour apporter de la valeur aux données. Le filtrage, le nettoyage et l'élimination réduisent le volume de données transféré vers le cache suivant et suppriment les données inappropriées ou corrompues. Pour être honnête, il existe un courant de pensée qui n'applique la logique de nettoyage qu'au moment de l'exécution de l'analyse, car il considère que le nettoyage peut supprimer des « pistes intéressantes ». Le « partage » consiste à placer les attributs de l'entreprise gérés à un niveau supérieur dans des entités majeures, comme les clients, les produits et les dates. L'existence de ces attributs permet d'établir des jonctions de grande valeur entre les différents domaines d'application. Cette étape pourrait simplement s'intituler « intégration ». Le diagnostic permet d'ajouter des attributs intéressants aux données, y compris des marqueurs de confiance et des identifiants textuels représentant des clusters de comportement identifiés par un professionnel de l'exploration des données. La découverte des données et le profilage aident à identifier les domaines de données, les relations, les balises de métadonnées utiles pour la recherche, les données sensibles et les problèmes de qualité des données.

Implémentez les retours des données, surtout à partir du data warehouse d'entreprise, aux caches situés en amont sur l'autoroute de l'information. Les dimensions de référence administrées au plus haut niveau du data warehouse d'entreprise, comme les clients, les produits et les dates, doivent être reconnectées aux données des premiers caches. Dans l'idéal, tout ce qu'il vous faut, ce sont des clés uniques et durables pour ces entités dans tous les caches. Le corollaire ici, c'est que chaque étape ETL d'un cache à un autre doit en premier lieu (première priorité) remplacer les clés propriétaires par des clés durables uniques afin que l'analyse de chaque cache puisse profiter du riche contenu en amont via une simple jonction à la clé durable. Cette étape ETL peut-elle être effectuée même lors du transfert de données sources brutes dans le cache en temps réel en moins d'une seconde ?
Peut-être…

Les données de dimension ne sont pas les seules à parcourir le chemin du retour vers la source. Les données dérivées des tables de faits, comme les récapitulatifs historiques et les résultats de l'exploration des données complexes peuvent être regroupées sous forme d'indicateurs simples ou de totaux agrégés, puis transférées aux premiers caches de l'autoroute de l'information. Pour finir, les liens de référence comme les clés ou les codes utiles peuvent être incorporés aux caches de données de faible latence afin de permettre à l'analyste de les lier à d'autres données pertinentes en un simple clic.

Passez aux analyses des données en streaming dans les flux de données sélectionnés. L'un des points de vue qui va de pair avec la faible latence, c'est l'envie de lancer une analyse sérieuse des données dès qu'elles entrent dans le flux, et si possible bien avant que le processus de transfert ne prenne fin. Les systèmes d'analyse en streaming suscitent un intérêt grandissant car ils permettent aux requêtes SQL de traiter les données à mesure qu'elles accèdent au système. Dans certains cas, lorsque les résultats d'une requête en streaming dépassent un seuil fixé, il est possible d'interrompre l'analyse avant que le processus n'aille à son terme. Le langage de requête continue CQL a permis de réaliser d'importants progrès au niveau de la définition des critères de traitement des données en streaming, avec notamment une sémantique intelligente permettant de déplacer les fenêtres temporelles de façon dynamique sur les données en streaming. Recherchez les extensions du langage CQL et les capacités d'interrogation des données en streaming dans les programmes de chargement pour les jeux de données SGBDR et HDFS déployés. Dans une mise en œuvre idéale, l'analyse des données en streaming aurait lieu pendant le chargement des données au rythme de plusieurs Go/s.

Implémentez des limites d'évolutivité larges de manière à éviter les « accidents dus aux limites du système ». Au tout début de la programmation informatique, lorsque les machines disposaient de disques durs et de mémoires réelles d'une capacité ridicule, les pannes système étaient monnaie courante et la hantise des développeurs d'applications. Lorsqu'une application a épuisé l'espace disque ou la mémoire réelle, le développeur doit se résoudre à élaborer des mesures qui requièrent généralement un gros effort de programmation qui n'ajoute rien au contenu principal de l'application. Les pannes dues aux limites des systèmes ont plus ou moins été éliminées pour les applications de bases de données classiques,

mais avec le big data, ce problème resurgit. Hadoop est une architecture qui réduit considérablement les problèmes d'évolutivité liés à la programmation car il est possible, dans la plupart des cas, d'ajouter indéfiniment du matériel courant. Bien entendu, ce matériel a besoin d'être configuré, branché et de disposer de connexions réseau à large bande passante. L'objectif est d'anticiper au maximum pour pouvoir adapter les systèmes aux volumes et débits colossaux.

Effectuez des essais de big data sur un cloud public puis passez à un cloud privé. L'avantage d'un cloud public, c'est qu'il peut être approvisionné et mis à l'échelle instantanément, à l'instar d'Amazon EMR et Google BigQuery. Dans ces deux cas, où la sensibilité des données autorise un prototypage rapide des transferts, cette fonction peut se révéler très utile. Rappelez-vous seulement de ne pas laisser d'importants jeux de données en ligne sur un cloud public le week-end, lorsque les programmeurs sont absents ! Cependant, si vous essayez d'exploiter les données localement avec des processus MapReduce respectant les racks, il se peut que vous ne puissiez pas utiliser un service de cloud public car il risque de ne pas vous offrir le contrôle souhaité sur le stockage des données.

Attendez-vous à des améliorations de performances multipliées par 10 ou 100 dans le temps en acceptant le changement de paradigme pour l'analyse à des vitesses très élevées. L'ouverture du marché du big data a encouragé la mise au point de solutions spécialement codées pour des types d'analyses précis. C'est à la fois un bien et un mal. Une fois libérés du contrôle exercé par l'optimiseur SGBDR et sa boucle interne, les développeurs peuvent implémenter des solutions ponctuelles qui sont 100 fois plus rapides que les techniques standards. Par exemple, des progrès phénoménaux ont été accomplis pour résoudre le problème de la « grande jointure » où une dimension d'un milliard de lignes est associée à une table de faits de mille milliards de lignes. À titre d'exemple, consultez l'approche de Yahoo en matière de gestion des jointures des jeux de données volumineux, ainsi que les projets Dremel et BigQuery de Google. Le problème, c'est que ces solutions ponctuelles ne font pas encore partie d'une architecture unique et unifiée.

La visualisation des jeux de données représente l'un des enjeux majeurs du big data. « Survoler » un pétaoctet de données nécessite des performances spectaculaires ! La visualisation du big data est un tout nouveau domaine de développement passionnant qui autorise l'analyse et la découverte de fonctions inattendues ainsi que le profilage des données. Autre application intéressante qui exige des performances élevées : « le zoom sémantique sans agrégation préalable ». L'analyste progresse depuis un niveau hautement agrégé vers des niveaux toujours plus détaillés au sein de données non structurées ou semi-structurées, à la manière d'un zoom sur une carte.

Retenons de cette pratique que les gains de performance multipliés par 10 ou 100 vont entraîner des progrès au niveau de notre capacité à consommer et analyser le big data, et que nous devons nous préparer à ajouter ces développements à notre suite d'outils.

Séparez les charges de travail analytiques du big data des data warehouses d'entreprise traditionnels de manière à préserver leurs niveaux de service contractuels. Si vos données de big data sont hébergées sur Hadoop, alors elles disposent des ressources suffisantes pour ne pas chercher à utiliser celles du data

warehouse d'entreprise basé sur un SGBDR traditionnel. Toutefois, soyez prudent si vous exécutez les analyses de big data sur la machine dédiée au data warehouse d'entreprise, car les exigences du big data évoluent rapidement et inexorablement dans une seule direction, à savoir qu'il faut toujours plus de ressources de traitement.

Exploitez les capacités uniques des analyses dans les bases de données. Les principaux acteurs des systèmes SGBDR investissent tous massivement dans les analyses à l'intérieur des bases de données. Une fois que vous avez chargé les données dans les tables relationnelles, vous pouvez combiner SQL avec les extensions d'analyse de plusieurs façons. Parmi les récents développements importants au niveau des bases de données, on note l'acquisition de Netezza et SPSS par IBM, l'intégration de SAS dans Teradata et Greenplum, Exadata R Enterprise et la syntaxe PostgreSQL d'Oracle pour la programmation des analyses, ainsi que d'autres fonctions secondaires relatives à la boucle interne des bases de données. Toutes ces options permettent la mise à disposition des bibliothèques testées composées de centaines de routines d'analyse. Certaines plates-formes d'intégration des données proposent l'optimisation Pushdown pour exploiter les analyses dans les bases de données dans le cadre d'un flux de données ou d'un processus ETL.

Meilleures pratiques de modélisation des données appliquées au Big Data

Les pratiques suivantes concernent les structures logiques et physiques des données.

Pensez en « dimensions » : divisez le monde en dimensions et en faits. Les utilisateurs métiers trouvent le concept de dimension à la fois naturel et évident. Quel que soit le format des données, il est toujours possible de trouver les entités de base qui y sont associées, comme les clients, produits, services, emplacements ou dates. Concernant la meilleure pratique suivante, nous allons étudier comment, avec un peu de discipline, il est possible d'utiliser les dimensions pour intégrer les sources de données. Mais avant d'en arriver là, nous devons identifier les dimensions de chaque source de données et les associer à chaque niveau inférieur d'observation de données élémentaires. Ce processus, que l'on pourrait qualifier de « dimensionnement », est parfait pour les analyses des big data. Par exemple, on pourrait penser que le simple tweet « Ouah ! C'est génial ! » ne contient aucun élément susceptible d'être dimensionné. Cependant, avec un peu d'analyse, il est possible de connaître l'auteur (ou client, citoyen, patient), son emplacement, le produit (ou service, contrat, événement) dont il est question, les conditions du marché, le fournisseur, le temps, le groupe démographique, la session, l'événement déclencheur, le résultat final, etc. Certaines formes de dimensionnement automatiques sont nécessaires pour conserver de l'avance sur la haute vélocité des flux de données. Comme nous le verrons dans une meilleure pratique plus loin dans ce document, il est souhaitable que les données entrantes soient entièrement dimensionnées dès la phase d'extraction.

Intégrez des sources de données distinctes aux dimensions partagées. Les dimensions partagées représentent le lien qui réunit les sources de données distinctes et permet de les combiner dans une analyse unique. Les dimensions partagées sont probablement la meilleure pratique la plus puissante de l'environnement du data warehouse d'entreprise conventionnel dont le big data doit hériter.

L'intérêt des dimensions partagées repose sur la présence d'un ou de plusieurs attributs d'entreprise (champs) dans les versions des dimensions associées à des sources de données séparées. Par exemple, chaque processus orienté client d'une entreprise est assorti d'une variation de la dimension client. Ces variations de la dimension client peuvent avoir des clés, des définitions de champ voire une granularité différentes. Et même dans les pires cas d'incompatibilité des données, il est possible de définir un ou plusieurs attributs d'entreprise et de les intégrer dans toutes les variations de la dimension client. Par exemple, une catégorie « démographie » est un choix plausible par rapport aux clients. Ces dimensions partagées pourraient être associées à pratiquement à toutes les dimensions client, même celles situées aux plus hauts niveaux d'agrégation. Ensuite, les analyses portant sur cette catégorie « démographie » peuvent s'étendre à toutes les sources de données prises en compte via un simple processus tri-fusion, une fois les requêtes exécutées sur les différentes sources de données. Mieux encore, la phase d'introduction des attributs d'entreprise dans les bases de données distinctes peut s'effectuer progressivement, avec flexibilité et sans interruption, comme expliqué en

détail dans mon livre blanc parrainé par Informatica sur ce thème. Toutes les applications d'analyse existantes continueront de fonctionner à mesure que le contenu de la dimension partagée sera déployé.

Ancrez toutes les dimensions au moyen de clés de substitution durables. S'il y a bien une leçon à retenir de l'environnement du data warehouse d'entreprise, c'est de ne pas ancrer les entités majeures (clients, produits, dates) avec les « clés naturelles » définies par une application spécifique. Ces clés naturelles sont en réalité un piège et ne fonctionnent pas dans le monde réel. Elles ne sont pas compatibles entre applications et mal administrées. Pour chaque source de données, la première étape consiste à augmenter la clé naturelle provenant d'une source à l'aide d'une clé de substitution durable. « Durable » signifie qu'aucune règle métier ne peut modifier une clé. Les clés durables relèvent du domaine informatique, et non de la source de données. Par « substitution », je veux dire que les clés elles-mêmes sont de simples entiers affectés en séquence ou générés par un algorithme de hachage puissant qui garantit leur caractère unique. Une clé de substitution isolée ne possède pas de contenu applicatif. C'est un simple identifiant.

Le monde du big data fourmille de dimensions évidentes qui doivent posséder des clés de substitution durables. Plus haut dans ce document, lorsque nous avons parlé du chemin de retour des données le long de l'autoroute, nous nous sommes basés sur la présence de clés de substitution durables pour que ce processus fonctionne. Nous avons également déclaré que la première priorité de chaque opération d'extraction des données d'une source brute consistait à incorporer les clés de substitution durables dans les dimensions appropriées.

Attendez-vous à devoir intégrer des données structurées et non structurées. Le big data complique considérablement le défi de l'intégration. La majeure partie du big data ne sera jamais placée dans une base de données relationnelle, mais demeurera sans doute dans un système Hadoop ou dans un grid. Néanmoins, une fois que nous sommes équipés de dimensions partagées et de clés de substitution durables, nous pouvons combiner toutes les formes de données dans des analyses individuelles. Prenons l'exemple d'une étude médicale : on sélectionne un groupe de patients possédant des attributs de santé et des données démographiques spécifiques. Il est alors possible de combiner leurs données du data warehouse d'entreprise classiques avec des données d'images (photographies, radios, ECG), des données de texte libre (notes des médecins), des impressions évoquées sur les réseaux sociaux (avis relatifs au traitement) et des liens à des groupes comparables (patients dans une situation similaire).

Sur le plan architectural, cette phase d'intégration doit avoir lieu lors de l'exécution des requêtes et non lors du chargement des données ou de la création de la structure. Pour réaliser l'intégration dans des conditions optimales, pensez à la virtualisation : les jeux de données intégrés apparaissent sous forme de tables physiques mais sont en réalité des spécifications similaires aux vues relationnelles dans lesquelles les sources de données distinctes sont unies au moment des requêtes. Si vous n'avez pas recours à la virtualisation des données, alors la couche de BI finale doit accomplir l'intégration.

Établissez un suivi des écarts de temps dans les dimensions à évolution lente. Suivre les écarts de temps dans les dimensions est une meilleure pratique ancienne et louable empruntée au monde des data warehouses d'entreprise. En résumé, elle contribue à notre promesse d'effectuer le suivi de l'historique avec minutie. Il n'est pas acceptable d'associer le profil actuel d'un client (ou citoyen, patient, étudiant) à un historique obsolète. Dans le pire des cas, le profil actuel est complètement faux lorsqu'il est appliqué à un historique trop ancien. Le traitement des dimensions à évolution lente peut prendre trois formes. La dimension à évolution lente de type 1 écrase le profil lorsqu'une modification a lieu, d'où la perte de l'historique. Cette action est possible lorsqu'il s'agit de corriger une erreur de données. La dimension à évolution lente de type 2, la technique la plus utilisée, génère un enregistrement de dimension modifié lorsqu'un changement a lieu. Lors de la génération du nouvel enregistrement de la dimension, la dimension à évolution lente de type 2 requiert la conservation de la clé de substitution durable qui lie le nouvel enregistrement aux anciens. Par ailleurs, il est également nécessaire de générer une clé principale unique pour l'instantané du membre de la dimension. Comme pour les dimensions partagées, ce processus a été décrit et étudié dans ses moindres détails. Pour finir, la dimension à évolution lente de type 3, qui n'est pas aussi courante que les deux premières, traite les situations où une « réalité alternative » est définie et coexiste avec la réalité « actuelle ». Merci de consulter mes articles d'introduction aux dimensions à évolution lente ainsi que les manuels où j'aborde dans le détail le thème des dimensions à évolution lente. En ce que concerne le big data, le point à retenir est qu'il est tout aussi important d'associer le profil actuel approprié d'une entité majeure à l'historique, comme c'est le cas dans l'univers du data warehousing d'entreprise.

Prenez l'habitude ne pas déclarer les structures de données avant le moment de l'analyse. L'un des atouts du big data, c'est qu'il est possible de différer la déclaration des structures de données, généralement exécutée au moment du chargement dans Hadoop ou dans un grid. Voilà qui procure de sérieux avantages... En premier lieu, il n'est pas nécessaire de comprendre les structures de données au moment du chargement. En effet, le contenu des données peut être tellement variable que le fait de posséder une structure unique serait inapproprié ou vous obligerait à modifier certaines données pour les adapter à la structure. Si vous chargez des données dans Hadoop, par exemple, sans déclarer leur structure, vous évitez une étape qui mobilise beaucoup de ressources. En dernier lieu, les différents analystes peuvent très bien « voir » les mêmes données différemment. Naturellement, cet aspect peut être pénalisant dans certains cas, car il peut s'avérer difficile, voire impossible, d'indexer les données sans structure déclarée pour les

rendre rapidement accessibles comme dans un SGBDR. Toutefois, la plupart des algorithmes d'analyse du big data traitent des jeux entiers de données sans que l'on attende un filtrage précis des sous-ensembles de données.

Cette meilleure pratique est en désaccord avec les méthodologies SGBDR traditionnelles qui mettent l'accent sur la modélisation minutieuse des données avant leur chargement. Mais ce conflit n'est pas insoluble. Pour les données destinées à un SGBDR, le transfert depuis un environnement Hadoop ou grid, mais aussi depuis une structure de paire nom-valeur, vers les colonnes nommées d'un SGBDR peut être considéré comme une étape ETL importante.

Créez la technologie autour des sources de données constituées de paires nom-valeur. Les sources du big data regorgent de surprises. Dans bien des cas, vous découvrez du contenu inattendu ou des données non documentées que vous devez quand même charger au rythme de plusieurs gigaoctets par seconde. Comment faire ? Tout simplement charger les données sous forme de paires nom-valeur. Par exemple, imaginez que des actifs financiers soient déclarés de la façon suivante, pour le moins inattendue : « timbre postal rare d'une valeur de 10 000 euros ». Dans un jeu de données comportant des paires nom-valeur, cette association serait chargée sans problème même si vous n'avez jamais vu la déclaration « timbre rare » auparavant et que vous ne sauriez quoi en faire lors du chargement. Bien entendu, cette pratique s'intègre parfaitement à la pratique précédente, c'est-à-dire le report de la déclaration des structures de données après leur chargement.

La structure de programmation MapReduce requiert que les données soient présentées sous la forme de paires nom-valeur, ce qui semble logique compte tenu de la présentation « générale » que peut adopter le big data.

Utilisez la virtualisation des données pour permettre le prototypage et les altérations de schéma rapides. La virtualisation des données est une technique idéale pour déclarer différentes structures de données logiques sur des données physiques sous-jacentes. Les définitions de vue standards dans SQL sont représentatives de la virtualisation des données. En théorie, la virtualisation des données peut permettre de présenter une source de données sous n'importe quel format recherché par l'analyste. Mais si la virtualisation supprime les coûts informatiques liés à l'exécution, elle engendre en contrepartie des coûts ETL liés à la création de tables physiques avant l'exécution. La virtualisation des données est parfaite pour créer des prototypes de structures de données et les modifier rapidement ou offrir d'autres solutions. La meilleure stratégie de virtualisation des données consiste à attendre la matérialisation des schémas virtuels une fois qu'ils ont été testés et examinés avant que les analystes ne souhaitent obtenir des améliorations des performances des tables physiques réelles.

Meilleures pratiques de gouvernance des données appliquées au Big Data

Les meilleures pratiques suivantes s'appliquent à la gestion de vos données en tant que ressources d'entreprise précieuses.

Rien n'a plus d'importance que la gouvernance du big data. Maintenant que nous avons capté votre attention, notez que la gouvernance du big data doit être une approche complète de tout votre écosystème de données et non seulemetn une solution ponctuelle pour le big data. La gouvernance des données appliquée au big data doit être une extension de votre approche de la gouvernance de toutes vos données d'entreprise. Nous avons présenté une étude expliquant comment améliorer le big data par l'intégration d'autres formes de données existantes, comme les données de votre data warehouse d'entreprise. Mais une intégration réussie alors que la gouvernance des données est établie (ou ignorée) pour le big data de manière isolée comporte des risques. La gouvernance des données englobe, au minimum, la confidentialité, la sécurité, la conformité, la qualité des données, la gestion des métadonnées, la gestion des données de référence et la création du glossaire d'entreprise qui expose les définitions et le contexte à la communauté métier. Il s'agit d'une liste imposante de responsabilités et de compétences que le département informatique ne doit pas tenter de définir sans un soutien important de la part de la direction, qui doit elle-même comprendre la portée de l'effort et prendre en charge la coopération interfonctionnelle requise.

Dimensionnez les données avant d'appliquer la gouvernance. Voici un défi intéressant proposé par le big data : vous devez appliquer les principes de gouvernance des données même lorsque vous ne savez pas à quoi vous attendre quant à leur contenu. Vous recevez les données à une cadence pouvant atteindre plusieurs gigaoctets/seconde, généralement sous la forme de paires nom-valeur et de contenu inconnu. Pour les classer de manière optimale en fonction de vos responsabilités en matière de gouvernance des données, il est recommandé d'en dimensionner autant que possible et dès que possible dans le pipeline de données. Analysez-les, rapprochez-les et appliquez une résolution d'identité à la volée. Nous avons déjà évoqué ce point lors de la discussion sur les avantages de l'intégration des données, mais nous plaidons ici contre l'utilisation des données avant cette étape de dimensionnement.

Si l'analyse des jeux de données inclut l'identification d'informations concernant des personnes ou des entreprises, alors la confidentialité est l'aspect le plus important de la gouvernance lors de l'incorporation de ce type d'informations. Même si chaque facette de la gouvernance des données revêt une importance capitale, la confidentialité est celle qui comporte le plus de responsabilités et de risques pour l'entreprise. La compromission de la vie privée de personnes ou de groupes de personnes peut ternir votre réputation, fragiliser la confiance du marché envers votre entreprise, mais aussi vous exposer à des poursuites civiles et pénales.

Ces compromissions peuvent également représenter un obstacle au partage de jeux de données pourtant riches entre entreprises, institutions, tiers et même au sein des entreprises, ce qui limite considérablement l'intérêt du big data dans des secteurs tels que la santé, l'éducation et juridique. La prolifération de données personnelles auxquelles nous avons accès menace d'émousser nos sens et d'endormir notre vigilance. Pour la plupart des formes d'analyse, les détails personnels doivent au minimum être masqués et les données suffisamment agrégées pour ne pas permettre l'identification des personnes. Notez (au moment de la rédaction de ce document) qu'il est important d'apporter une attention particulière au stockage des données sensibles dans Hadoop, car une fois les données écrites dans Hadoop, ce dernier ne gère pas très bien les mises à jour. De ce fait, les données doivent être masquées ou chiffrées lors de l'écriture (masquage persistant des données) ou lors de la lecture (masquage dynamique des données).

N'excluez pas la gouvernance des données dans votre hâte d'utiliser le Big Data. Même dans vos projets de prototypes exploratoires du big data, établissez une liste d'éléments à vérifier à mesure que vous progressez. Vous ne voulez pas d'une bureaucratie inefficace, mais vous voudrez certainement obtenir une bureaucratie agile ! Cette liste, que vous devez veiller à mettre à jour, doit :

- Vérifier qu'il existe une vision et une étude de cas indiquant les orientations et les priorités.
- Identifier le rôle des intervenants, notamment les gestionnaires de données, les sponsors, les initiateurs de programmes et les utilisateurs.
- Vérifier l'implication de l'entreprise et les comités de partenariat/direction inter-entreprise en vue de la prise en charge de la remontée des informations et de leur hiérarchisation.
- Déterminer les outils et l'architecture requis et existants qui prendront en charge le cycle de vie du big data géré.
- Incorporer quelques notions de politiques d'utilisation des données et des normes de qualité des données.
- Inclure la gestion flexible des modifications organisationnelles pour tous les autres points de cette liste.
- Mesurer les résultats, la valeur opérationnelle ainsi que le retour sur investissement pour l'entreprise.
- Évaluer et influencer les processus dépendants en amont et en aval afin de minimiser le dilemme toujours présent de la qualité des données.

Résumé

Le big data s'accompagne d'une pléiade de changements et d'opportunités pour l'informatique et d'un tout nouvel ensemble de règles. Après une dizaine d'années de mise en œuvre, de nombreuses bonnes pratiques ont vu le jour. La plupart d'entre elles sont le prolongement de l'univers du data warehouse d'entreprise/de la BI et quelques autres représentent de nouvelles méthodes d'appréhender les données et la mission de l'informatique. Reconnaître que cette mission s'est étendue est à la fois logique et en quelque sorte dépassé. Actuellement, avec la prolifération des canaux de collecte des données, les nouveaux types de données et les nouvelles opportunités d'analyse, la liste des meilleures pratiques va continuer de s'allonger de manière intéressante.

Remerciements

Je remercie vivement les experts en big data suivants avec lesquels j'ai eu le privilège de m'entretenir au cours de la préparation de ce livre blanc. Leurs noms sont indiqués ci-dessous suivant l'ordre alphabétique des entreprises :

EMC : Bill Schmarzo

Fannie Mae : Javed Zaidi

Federal Aviation Administration : Wayne Larson et ses collègues

Informatica : John Haddad, Robert Karel

JP Morgan Chase : Tom Savarese

Kimball Group : Margy Ross

Merck : John Maslanski

Rackspace : Pete Peterson

VertiCloud : Raymie Stata

Yahoo : George Goldenberg

Références

Livre blanc Kimball sponsorisé par Informatica sur le Big Data et Hadoop : http://vip.informatica.com/?elqPURLPage=8808

Livre blanc Kimball sponsorisé par Informatica sur les dimensions partagées et l'intégration des données : http://vip.informatica.com/?elqPURLPage=807

Articles Kimball de présentation des dimensions à évolution lente : http://www.kimballgroup.com/2008/08/21/slowly-changing-dimensions/

http://www.kimballgroup.com/2008/09/22/slowly-changing-dimensions-part-2/

L'approche Yahoo du défi de la grande jonction : http://www.slideshare.net/Hadoop_Summit/yahoo-display-advertising-attribution

Présentation du zoom sémantique sans agrégation préalable : http://www.slideshare.net/Hadoop_Summit/empowering-semantic-zooming-with-hadoop-and-hbase