talend*
*open integration solutions

White Paper

Les quatre piliers d'une solution de gestion des Big Data

Table des Matières

Introduction

Les gros volumes de données (« Big Data ») représentent une évolution significative des modèles technologiques de l'entreprise. Le phénomène Big Data change radicalement les modalités de gestion des données puisqu'il introduit de nouvelles problématiques concernant la volumétrie, la vitesse de transfert et le type de données. Il permet d'adapter les technologies et les stratégies d'entreprise en fournissant des informations critiques pour des actions ciblées et optimisées, et offre la possibilité d'accéder à de nouvelles opportunités d'affaires et de mieux contrôler les risques inhérents. Par ailleurs, il est probable qu'il transformera l'entreprise moderne telle que nous la connaissons aujourd'hui. Voici les conclusions évidentes que nous pouvons en tirer :

- Le Big Data aborde les besoins réels du marché en s'appuyant sur une nouvelle avancée technologique.
- Alors que la plupart des entreprises sont en phase de recherche, de nombreux modèles d'utilisation du Big Data ont fait leur apparition.
- Si l'intégration de données est essentielle au traitement des Big Data, la qualité et la gouvernance de données n'en restent pas moins des préoccupations majeures.
- Les projets de Big Data quitteront la sphère de l'expérimentation pour devenir un atout stratégique pour l'entreprise.
- Des outils de développement sont nécessaires pour augmenter l'adoption de ces nouvelles technologies et ainsi réduire le recours obligatoire aux développeurs très qualifiés. Tous les principaux vendeurs d'infrastructures et de bases de données commencent à lancer des solutions de Big Data sur le marché.

Big Data : un terme très vaste

"Big Data" est un concept très difficile à définir avec précision, puisque la notion même de "big" en termes de volumétrie des données varie d'une entreprise à l'autre. Il n'est *pas* défini par un ensemble de technologies, bien au contraire, il *définit* une catégorie de techniques et de technologies. Il s'agit d'un domaine émergent et alors que nous cherchons à apprendre comment implémenter ce nouveau paradigme et en exploiter la valeur, la définition se transforme. Pourtant, si celle-ci peut s'avérer ambiguë, de nombreux experts sont convaincus que des secteurs entiers de l'industrie et des marchés seront atteints et d'autres créés, au fur et à mesure que ces capacités permettront la réalisation de nouveaux produits et fonctionnalités qui étaient inimaginables auparavant.

Le "Big" Data

Comme l'expression l'indique, le Big Data se caractérise par la taille ou la volumétrie des informations. Mais d'autres attributs, notamment la **vitesse** et le **type** de données, sont aussi à considérer. En ce qui concerne le **type**, le Big Data est souvent rattaché à du contenu non structuré ou semi-structuré, ce qui peut représenter un défi pour les environnements classiques de stockage relationnel et de calcul. Les données non structurées et semi-structurées sont partout : contenu web, posts twitter ou commentaires client en format libre. Par **vitesse** on entend la rapidité avec laquelle les informations sont créées. Grâce à ces nouvelles technologies, il est maintenant possible d'analyser et d'utiliser l'importante masse de données fournie par les fichiers log des sites web, l'analyse d'opinions des réseaux sociaux, et même les vidéos en streaming et les capteurs environnementaux. Nous pouvons ainsi tirer parti d'une vision stratégique impossible jusqu'à ce jour.

Afin d'avoir un aperçu de la complexité engendrée par la volumétrie, la vitesse et le type de données, il suffit de considérer les cas suivants :

- Walmart traite chaque heure plus d'un million de transactions clients, qui sont importées vers des bases de données dont le contenu est estimé à plus de 2,5 petabytes - à savoir 167 fois l'équivalent des informations réunies dans tous les livres de la Bibliothèque du Congrès américain.
- Facebook gère 40 milliards de photos à partir de sa base d'utilisateurs.
- Décoder le génome humain a demandé initialement 10 ans de travail ; cette tâche peut maintenant être effectuée en une semaine.
- La distribution Hadoop par Hortonworks gère à présent plus de 42,000 machines Yahoo! traitant des millions de requêtes par jour.

De tels exemples de technologies de pointe deviennent de plus en plus répandus au fur et à mesure que les entreprises comprennent à quel point ces immenses magasins de données contiennent des informations inestimables pour leur activité.

La technologie "Big Data"

Afin de saisir l'impact de ce nouveau paradigme, il est important d'avoir des connaissances de base sur les nouveaux concepts, ainsi que sur les termes et les technologies clé, qui définissent le Big Data. Au cœur de cette révolution, une architecture connue sous le nom de MapReduce : elle consiste en un puissant environnement massivement parallèle qui exécute des fonctions avancées en très peu de temps.
Introduit par Google en 2004, MapReduce permet à un programmeur d'exécuter une transformation de données, qui sont ensuite déléguées et traitées par une architecture « cluster » pouvant inclure des milliers d'ordinateurs qui opèrent simultanément. À la base, il s'articule en deux étapes : "map", où les problèmes sont découpés et distribués à plusieurs

serveurs parallèles, et "reduce", où les réponses sont consolidées à partir de chaque map et résolvent problème initial.

De nombreuses technologies Big Data, telles que Hadoop, Pig et Hive, sont disponibles dans l'offre open source. Celle-ci propose, contrairement aux logiciels propriétaires, de nombreux avantages : innovation continue, réduction des coûts, interopérabilité et développement basé sur des standards

Une description plus approfondie du fonctionnement de l'architecture MapReduce, ainsi qu'un glossaire des technologies Big Data, sont reportés en annexe de ce document.

Le grand changement de paradigme

Les technologies Big Data ont déjà révolutionné notre mode de vie. Si Facebook, Groupon, Twitter, Zynga et d'autres nouveaux et nombreux modèles métiers existent, c'est grâce à l'avance générée par le Big Data. Il s'agit là d'un changement de paradigmes technologiques qui pourrait avoir des conséquences plus marquantes que la commercialisation d'Internet à la fin des années 90. Il impactera des secteurs entiers de l'industrie et des marchés au fur et à mesure que nous apprendrons à utiliser ces capacités pour fournir non seulement un meilleur résultat et des nouvelles fonctionnalités dans les produits proposés, mais aussi la création de nouvelles solutions auxquelles on ne croyait pas il y a peu.

Considérons à titre d'exemple la vue unique du client telle qu'obtenue au moyen d'outils MDM. Les solutions actuelles se basent sur une base relationnelle quelque peu statique pour maintenir les données et doivent exécuter un algorithme par lots pour créer cette image globale. Les limites actuelles de performance et de stockage réduisent ces solutions à l'utilisation d'un ensemble explicite de données. Hadoop élimine ces restrictions de telle sorte qu'une vue unique du client peut être générée à tout moment et inclure davantage d'informations, comme par exemple les données relationnelles. *Comment pourrions-nous utiliser l'analyse des sentiments exprimés dans les réseaux sociaux pour élargir la vue du client ?*

Cette avancée risque de perturber de nombreux marchés existants. Pensons aux ERP et aux entrepôts de données, où le Big Data joue un rôle important pour le futur de la gestion des data warehouse et des produits analytiques. *Et si nous utilisions les technologies Big Data pour remplacer une base de données opérationnelle ?* C'est une réflexion radicale mais proche de la réalité, puisque les outils open source se basant sur le Big Data peuvent servir à développer, et, dans une certaine mesure, relever quelques-unes de ces fonctions, offrant ainsi une nouvelle perspective sur les modalités actuelles de gestion des données. Nous assistons aujourd'hui aux prémices d'un immense changement technologique qui engendrera un immense changement sociétaire.

Le Big Data transforme tout.

Big Datal : évolution des cas d'usage

Le Big Data est un phénomène émergent. Pourtant, des cas d'usage courants sont déjà connus et apportent dès à présent une valeur significative. En voici quelques exemples :

Moteur de recommandation

Depuis des années, des entreprises telles qu'Amazon, Facebook et Google utilisent des moteurs de recommandation pour filtrer et suggérer aux utilisateurs des produits, personnes et annonces, en fonction de l'analyse de leurs profils et des renseignements sur leur comportement en ligne. Les problèmes liés à l'analyse de ces volumes importants d'informations ont été parmi les premiers abordés par le Big Data et leur résolution a contribué à développer la technologie telle que nous la connaissons à l'heure actuelle.

Analyse de campagnes marketing

Un marketeur identifiera et touchera d'autant plus de cibles "granulaires" qu'il disposera d'informations. Le Big Data est utilisé pour analyser d'énormes quantités de données qui échappent aux solutions relationnelles classiques, de telle sorte que les spécialistes dans le domaine du marketing peuvent maintenant mieux repérer un public cible et associer les produits et services appropriés à un individu précis. Grâce au Big Data, ils étudient d'importants volumes d'informations à partir de nouvelles sources, comme le parcours de navigation ou les enregistrements des détails des appels, ce qui leur permet de mieux comprendre les tendances et les comportements d'achat des consommateurs.

Analyse de la fidélisation et de la perte de clientèle

Une augmentation du nombre de produits par client équivaut souvent à une diminution de la perte de clientèle, et de nombreuses sociétés entreprennent d'importants efforts pour améliorer les techniques de vente croisée et de montée en gamme. Toutefois, l'étude de la clientèle et des produits à travers les secteurs d'activité s'avère souvent difficile, puisque des formats hétérogènes de données et des problématiques de gouvernance limitent ces efforts. Certaines entreprises ont la possibilité de charger ces données dans un cluster Hadoop, afin d'effectuer des analyses à grande échelle pour identifier les tendances. Le résultat montre les clients susceptibles de partir à la concurrence ou, encore mieux, ceux qui vont probablement approfondir leur relation commerciale avec l'entreprise. Des mesures peuvent alors être adoptées pour reconquérir ou encourager ces clients selon le cas.

Analyse des graphes sociaux

Chaque réseau social ou communauté compte des utilisateurs ordinaires et des super-utilisateurs, et reconnaître ces derniers est une tâche difficile. Avec le Big Data, les données

issues des activités des réseaux sociaux sont explorées pour indiquer les membres exerçant une influence majeure sur le groupe. Ceci permet aux entreprises d'identifier les clients « les plus significatifs », qui ne sont pas forcément ceux utilisant l'offre de produits la plus vaste ou bénéficiant du budget conséquent, contrairement à la définition classique répandue dans le cadre de l'analyse décisionnelle.

Analyse des marchés des capitaux

Que nous recherchions de grands indicateurs économiques, ou des indicateurs de marché spécifiques ou bien encore des avis sur une entreprise donnée et ses actions, la richesse des informations à analyser est impressionnante tant en provenance des sources classiques que des nouveaux réseaux. Si l'analyse par mots clé de base et l'extraction d'entités sont utilisées depuis plusieurs années, l'association d'informations classiques et de sources inédites telles que Twitter et d'autres médias sociaux permettent d'accéder à un aperçu détaillé de l'opinion publique, pratiquement en temps réel. Aujourd'hui, la plupart des institutions financières se servent, à différents degrés, de l'analyse des sentiments pour mesurer la perception du public sur leur entreprise, sur le marché, ou sur l'économie en général.

Analyse prédictive

Afin de prévoir les changements économiques, les experts dans le domaine des marchés des capitaux confrontent d'un côté les algorithmes de corrélation avancés et calculs des probabilités, et, de l'autre, les données historiques et actuelles. Le volume important des archives d'informations sur les marchés ainsi que la vitesse exigée pour l'évaluation des nouveaux renseignements (par exemple : valorisations complexes d'instruments dérivés) font de l'analyse prédictive un problème majeur que le Big Data contribue à résoudre. En effet, grâce à la capacité à effectuer ce type de calculs plus rapidement, et avec du matériel informatique courant, le Big Data remplace de manière fiable l'approche relativement lente et coûteuse fondée sur les systèmes traditionnels.

Gestion des risques

Les entreprises dont la technologie se veut avancée et déterminée tentent de minimiser les menaces au moyen d'une gestion continue des risques et d'une analyse à large spectre des facteurs de risque, en croisant de vastes catégories de données. Par ailleurs, une demande de plus en plus pressante oblige à accélérer l'analyse des informations, malgré leur volume toujours croissant. Les technologies de Big Data s'imposent dans la résolution de ce problème : en effet, les calculs peuvent être effectués tout en accédant aux données. Qu'il s'agisse d'analyse croisée ou d'intégration d'informations sur les risques et les tendances financières, afin de rajuster les rendements et les bilans, il est nécessaire de fusionner, de permettre l'accès et de traiter à tout moment une quantité grandissante de données provenant des différents services autonomes de l'entreprise.

"Rogue trading"

Une analyse approfondie reliant les données comptables aux systèmes de repérage et de gestion des commandes peut fournir des informations stratégiques précieuses qui ne seraient pas disponibles avec les outils classiques. Afin de les identifier, une masse importante de données doit être traitée presque en temps réel à partir de sources multiples et hétérogènes. Cette fonction permettant de puissants calculs peut maintenant être effectuée par le biais des technologies Big Data.

Détection des fraudes

Mettre en rapport des données à partir de sources multiples et non reliées augmente la possibilité d'identifier des activités frauduleuses. Si, dans le cadre du Big Data, l'on relie par exemple des mouvements bancaires effectués en ligne, aux distributeurs automatiques, via smartphone, par carte de paiement, à l'analyse du comportement web retracé sur le site de la banque où ailleurs, la détection des fraudes en est améliorée.

Services bancaires de détail

Dans le domaine des services bancaires de détail, la capacité de déterminer avec précision le niveau de risque sur le profil d'un individu ou sur un prêt joue un rôle primordial dans la décision d'attribuer (ou de refuser) à un client certaines prestations. Comprendre correctement la situation protège non seulement la banque, mais satisfait aussi le client. Un accès à des informations exhaustives sur la clientèle permet aux banques de bénéficier de garanties et de visibilité afin de mieux cibler les offres de services. Il est aussi possible de prévoir les événements significatifs dans la vie du client, tel un mariage, la naissance d'un enfant, l'achat d'une maison, ce qui est un atout pour appuyer les activités de vente croisée et de montée en gamme.

Surveillance du réseau

Tous les types de réseaux, qu'il s'agisse de transports, de communications ou de protection policière, peuvent bénéficier d'une meilleure analyse, activité dans laquelle interviennent les technologies Big Data. Considérons par exemple le réseau local d'un bureau : grâce au Big Data, des volumes considérables d'informations sont acheminés depuis des serveurs, des périphériques et du matériel informatique divers. Les administrateurs peuvent ainsi surveiller l'activité du réseau et détecter des congestions et bien d'autres problèmes avant qu'ils n'aient un impact négatif sur la productivité.

Recherche et développement

Les entreprises qui disposent de services de recherche et développement importants, comme les établissements pharmaceutiques, se servent des technologies Big Data pour examiner

minutieusement d'énormes quantités d'informations texte et de données historiques afin d'accompagner la conception de nouveaux produits.

Les défis spécifiques du Big Data

Le Big Data représente certes une opportunité significative, mais pose aussi des défis spécifiques. Ces derniers incluent un ensemble relativement nouveau de technologies plutôt complexes à appréhender, dépourvues pour l'instant d'outils pour encourager leur adoption et leur développement. Par ailleurs, les ressources documentées sont peu nombreuses. En effet, les projets de Big Data, dans leur majorité, ne sont actuellement qu'au stade de l'ébauche, si bien qu'ils ne sont pas incorporés au cadre de gouvernance attendu pour la gestion de projets et de données à l'échelle de l'entreprise. Cette situation est cependant destinée à évoluer. Examinons maintenant plus en détail les challenges mentionnés.

Ressources limitées

La majorité des développeurs et des architectes qui "comprennent" le Big Data travaillent pour les créateurs de technologies de Big Data, à savoir Facebook, Google, Yahoo et Twitter, pour en citer quelques-uns, ou pour les nombreuses startups dans le domaine, comme Hortonworks, Cloudera et Mapr. Comme la technologie reste un peu difficile à maîtriser, le taux de disponibilité des ressources sur le Big Data s'en trouve restreint. A cette problématique s'ajoute le fait que dans ce marché naissant, les outils de support au développement et à l'implémentation des projets sont peu nombreux.

Faible qualité des données + Big Data = problèmes en vue

Selon l'objectif d'un projet de Big Data, une faible qualité des données risque d'avoir un impact important sur son efficacité. De même, dans le cadre du Big Data, des informations incomplètes ou incohérentes peuvent affecter le traitement de manière exponentielle. Au fur et à mesure que l'analyse basée sur le Big Data se répandra, la nécessité de maîtriser validation, standardisation, enrichissement et résolution des données s'amplifiera également. L'identification même des liens peut être considérée comme un souci de qualité des données à résoudre pour l'implémentation du Big Data.

Gouvernance du projet

À l'heure actuelle, les projets Big Data correspondent souvent à une demande non prioritaire de la part des directeurs de l'entreprise, qui souhaitent en savoir davantage sur le sujet. Le processus d'adoption en est à ses premières phases, de telle sorte que les entreprises cherchent pour la plupart à déterminer la valeur potentielle en mettant en place des équipes dédiées. Dans le territoire encore inexploré du Big Data, ces recherches sont conduites sans

aucune direction. Quoi qu'il en soit, comme toute exigence en matière d'administration des données de l'entreprise, elles finiront par aboutir au respect des standards et des normes pour l'organisation, le déploiement et le partage des artefacts.

Malgré les défis, la technologie est stable. Les perspectives de croissance et d'innovation sont vastes, puisque le cycle de vie complet de la gestion de données, y compris la qualité et la gouvernance, peut être transféré dans ce nouveau paradigme. Les technologies de Big Data sont porteuses d'un intérêt illimité et, une fois leur choix répandu, les individus talentueux combleront leur niveau de compétences pour l'appuyer.

Quatre facteurs déterminants pour une solution de gestion des Big Data

L'intégration est « le moteur » et la génération de code le « carburant ».

Pour relever les défis décrits au chapitre précédent, au moment où l'on entreprend la construction d'une solution de gestion des Big Data, il faut garder à l'esprit quatre éléments déterminants : intégration, manipulation, qualité et gestion/gouvernance du Big Data. Talend, un des leaders mondiaux des solutions open source, fournit ces fonctionnalités réunies dans un environnement intuitif dédié à l'administration de données qui simplifie le développement, le déploiement et la gouvernance du Big Data.

Intégration des Big Data

Importer les Big Data (volumes importants de fichiers log, données générées par les systèmes opérationnels, réseaux sociaux, capteurs et sources diverses) dans Hadoop via HDFS, HBase, Sqoop ou Hive est considéré comme une tâche délicate pour l'intégration opérationnelle des données. Talend offre une solution immédiate pour relier directement à ces technologies Big Data les ressources classiques, comme les bases de données, les applications et les serveurs de fichiers.

Talend propose un espace de travail ainsi qu'un ensemble intuitif d'outils graphiques permettant une interaction avec une source ou une cible Big Data, sans besoin d'apprendre ni d'écrire du code complexe. Une fois la connexion Big Data configurée et représentée graphiquement, le code sous-jacent est généré de façon automatique, étant ainsi disponible pour le déploiement en tant que service, exécutable ou bien comme job stand-alone. La palette complète de composants de Talend pour l'intégration de données (applications, base de données, services et, même, un centre de données de référence) est utilisée de telle sorte que le mouvement des données peut être orchestré à partir de n'importe quelle source, et vers tout type de cible. Pour terminer, les outils graphiques fournis par Talend simplifient la configuration de technologies NoSQL (par exemple Hive et HBase) et garantissent un accès aléatoire au Big Data, en temps réel, en lecture et écriture, et orienté colonnes.

Manipulation des Big Data

Dans le cadre du Big Data, une gamme d'outils permet aux développeurs de bénéficier d'une parallélisation afin d'opérer des transformations sur des volumes massifs de données. Des plates-formes telles qu'Apache Pig procurent un langage de script pour comparer, filtrer, évaluer et regrouper les données au sein d'une architecture cluster HDFS. Talend résume ces fonctions en une gamme d'outils à partir desquels les scripts sont définis dans un environnement graphique en tant que partie du flux de données, et peuvent être développés rapidement sans aucune connaissance préalable du langage sous-jacent.

Qualité et Big Data

Talend offre des fonctionnalités de qualité des données qui bénéficient de l'environnement massivement parallèle de Hadoop et mettent à disposition tâches et fonctionnalités explicites pour établir le profiling des doublons et les identifier en quelques instants (au lieu de quelques jours) parmi les immenses quantités d'informations stockées. Ceci s'impose comme le complément logique des solutions de qualité et d'intégration des données de l'entreprise, ainsi que de ses bonnes pratiques.

Gestion et gouvernance de projets Big Data

Pour la plupart, les premiers projets Big Data en cours ne sont pas encadrés par une structure de gestion définie, mais ceci est destiné à évoluer avec l'intégration progressive de tels projets dans le système global. Ce changement entraînera pour les entreprises la nécessité de créer des standards et des procédures, comme cela a été le cas dans le passé pour l'administration des données. Talend propose une gamme de fonctionalités pour la gestion de projets Big Data au moyen de laquelle les entreprises peuvent programmer, surveiller et déployer toute sorte de job, tout en utilisant un référentiel commun où les développeurs échangent et partagent métadonnées et artefacts. Par ailleurs, Talend simplifie la génération de codes tels que Hcatalog et Oozie.

Talend et Big Data : solutions disponibles aujourd'hui

L'approche open source de Talend, ainsi que sa plate-forme flexible d'intégration pour le Big Data permet aux utilisateurs de relier et d'analyser facilement des données provenant de systèmes disparates pour contribuer à piloter et à améliorer la performance de l'entreprise. Les composants Big Data de Talend s'intègrent aux offres des éditeurs les plus importants dans le secteur, notamment Cloudera, Hortonworks, Google, Greenplum, Mapr, Teradata et Vertica, positionnant Talend comme le leader dans le traitement des Big Data. La mission de Talend est de démocratiser ce marché comme cela a été le cas avec l'intégration et la qualité des données, le MDM (Master Data Management), l'ESB (Enterprise Service Bus) et la gestion des processus métier (BPM).

Talend propose trois produits Big Data :

- Talend Open Studio for Big Data
- Talend Enterprise Data Integration – Big Data Edition
- Talend Platform for Big Data

Talend Open Studio for Big Data

Talend Open Studio for Big Data est un outil de développement open source gratuit qui combine les composants Big Data de Talend pour Hadoop, Hbase, Hive, Hcatalog, Oozie, Sqoop et Pig avec la base Talend Open Studio for Data Integration. Il est délivré dans la communauté sous licence Apache. Il permet aussi le passage des anciens systèmes aux nouveaux, puisqu'il inclut des centaines de composants pour les environnements existants, comme SAP, Oracle, DB2, Teradata et bien d'autres. Une version beta est accessible sur www.talend.com.

Talend Enterprise Data Integration _ Big Data Edition

Talend Enterprise Data Integration – Big Data Edition est une extension du produit Talend Open Studio for Big Data avec support technique professionnel et fonctionnalités de niveau entreprise. Pour tirer parti d'une qualité avancée du Big Data ainsi que des fonctions d'administration de projet, l'upgrade vers cette version se fera à partir de Talend Enterprise Data Integration.

Talend Platform for Big Data

Cette plate-forme de Talend relève les défis de l'intégration, de la qualité et de la gouvernance des Big Data, et facilite l'importation, l'extraction et le traitement de volumes importants et

diversifiés de données, permettant ainsi une prise de décision plus rapide et efficace. Elle combine Talend Enterprise for Data Integration avec les composants pour la qualité des données d'entreprise, de telle sorte qu'il est possible d'identifier et de relier des éléments par le biais d'un environnement massivement parallèle tel que Hadoop.

Délivrée par Talend pour étendre sa plate-forme unifiée, Talend Platform for Big Data améliore la productivité à travers les domaines de la gestion de données en partageant un référentiel et des outils communs de code pour l'ordonnancement, la gestion des métadonnées, le traitement des informations et l'activation des services.

Pour plus d'informations sur les fonctionnalités de chaque version des produits Talend, nous vous invitons à visiter www.talend.com.

Conclusion

Les volumes de données massifs (« Big Data ») engendrent une évolution considérable des modèles technologiques de l'entreprise, évolution qui permet d'accéder à de nouvelles opportunités d'affaires et de mieux contrôler les risques inhérents.
Le Big Data représente certes une opportunité significative, mais pose aussi des défis spécifiques. Ces derniers incluent un ensemble relativement nouveau de technologies plutôt complexes à appréhender, dépourvues pour l'instant d'outils pour encourager leur adoption et leur développement. Par ailleurs, les ressources documentées sont peu nombreuses. L'approche open source de Talend, ainsi que sa plate-forme flexible d'intégration pour le Big Data relève ces défis, permettant aux utilisateurs de relier facilement et analyser des données provenant de systèmes disparates pour contribuer à piloter et améliorer la performance de l'entreprise.

Annexe : un aperçu des technologies

MapReduce comme framework

MapReduce permet à la technologie Big Data telle que fournie par exemple par Hadoop, de fonctionner. Le Hadoop Data File System (HDFS) utilise ces composants pour la gestion de la persistance, exécute des fonctions sur des données et trouve des résultats. Les bases de données NoSQL, notamment MongoDB et Cassandra, stockent et localisent les informations pour les services respectifs par le biais de ces fonctions. Hive utilise ce framework pour créer les fondations d'un entrepôt de données.

Fonctionnement de Hadoop

Hadoop a vu le jour parce que les approches existantes n'étaient pas adaptées au traitement de volumes massifs de données, et pour relever le défi de l'indexation quotidienne du web. Google a développé un paradigme appelé MapReduce en 2004, Yahoo! a mis en place Hadoop comme implémentation de MapReduce en 2005 et a fini par le lancer en tant que projet open source en 2007. À l'instar des autres systèmes opérationnels, Hadoop possède les structures de base nécessaires à effectuer des calculs : un système de fichiers, un langage de programmation, une méthode pour distribuer les programmes ainsi générés à un cluster, un mode pour obtenir les résultats et arriver à une consolidation unique de ces derniers, ce qui est le but.

Avec Hadoop, le Big Data est distribué en segments étalés sur une série de nœuds s'exécutant sur des périphériques de base. Au sein de cette structure, les données sont dupliquées à différents endroits afin de récupérer l'intégralité des informations en cas de panne. Les données ne sont pas organisées par rangs et par colonnes relationnelles, comme dans le cas de la gestion classique de la persistance, ce qui comporte une capacité à stocker du contenu structuré, semi-structuré et non structuré.

Il y quatre types de nœuds intervenant au sein de HDFS :

- Name Node: un facilitateur qui indique l'emplacement des données. Il reconnaît les nœuds disponibles et ceux qui ont échoué, et sait repérer les informations voulues dans le cluster.
- Secondary Node: un backup du Name Node.
- Job Tracker: coordonne le traitement des données utilisant MapReduce.
- Slave Nodes: stockent les données et exécutent les commandes du Job Tracker.

Le Job Tracker est le point d'entrée d'un "map job" ou d'un processus à appliquer aux données. Un map job, qui est généralement une requête écrite en Java, constitue la première étape du processus MapReduce. Le Job Tracker demande au Name Node d'identifier et trouver les informations nécessaires pour compléter le job. Une fois ces informations obtenues, il soumet la requête aux Name Nodes concernés. La caractéristique massivement parallèle de MapReduce réside dans le fait que tout traitement de données demandé a lieu dans chaque nœud nommé.

Quand chaque nœud a fini l'élaboration, il stocke les résultats. Le client commence alors un "reduce job". Les résultats sont ensuite consolidés pour déterminer la réponse à la requête initiale, puis accessibles sur le système de fichiers et prêts pour tout travail d'exploitation.

Pig

Le projet Apache Pig est un langage de programmation du flux de données de haut niveau ainsi qu'un framework d'exécution pour la création de programmes MapReduce dans Hadoop. Le langage pour cette plate-forme, connu sous le nom de Pig Latin, fournit une syntaxe abstraite de programmation pour la transformer en notation, ce qui apparente la programmation MapReduce à celle de SQL pour systèmes RDBMS. Avec les UDF (User Defined Functions), Pig Latin peut bénéficier d'extensions, que l'utilisateur peut écrire en Java, puis rappeler directement à partir du langage.

Hive

Apache Hive est une infrastructure d'entrepôt de données créée en complément de Hadoop (initialement par Facebook) pour fournir compression des données, lancement de requêtes ad hoc, et analyse de volumes massifs d'informations. Hive met à disposition un système pour structurer ces informations et les interroger par le biais d'un langage de type SQL appelé HiveQL, qui simplifie l'intégration avec les outils de BI et de visualisation.

HBase

HBase est une base de données non relationnelle installée sur le système de fichiers Hadoop (HDFS). Orientée colonnes, elle fournit un stockage en continu en cas de panne matérielle et un accès rapide à des quantités importantes de données éparses. HBase ajoute aussi des fonctionnalités transactionnelles à Hadoop, permettant aux utilisateurs d'effectuer des mises à jour, des insertions et des suppressions. Développé initialement par Facebook pour leurs services de messagerie, HBase est aussi largement utilisé par eBay.

HCatalog

HCatalog est un service de gestion des tables et du stockage créé avec Apache Hadoop qui permet le partage des données entre différents outils d'élaboration, tels que Pig, Map Reduce, Streaming, et Hive, sans requérir un changement de format de schéma.

Flume

Flume est un système d'agents peuplant un cluster Hadoop, et déployés à travers une infrastructure IT pour la collecte de données destinées à être stockées et traitées par Hadoop.

Oozie

Oozie coordonne les jobs écrits en différents langages tels que MapReduce, Pig et Hive. Il s'agit d'un système de gestion des workflows qui relie les jobs et qui permet d'en spécifier la commande et les dépendances.

Mahout

Mahout est une librairie d'exploration de données qui implémente des algorithmes très connus pour la modélisation statistique et le clustering.

Sqoop

Sqoop est un set d'outils pour l'intégration de données qui permet à des bases non-Hadoop d'échanger avec des bases et des entrepôts relationnels classiques.

NoSQL (not only SQL)

NoSQL désigne une vaste catégorie de systèmes de stockage de bases de données très différents de l'architecture classique des bases relationnelles (RDBMS). Ces technologies utilisent leur propre langage de requêtes et sont souvent construites sur des structures de programmation avancées pour relations de type clé/valeur, objets définis, méthodes tabulaires ou tuples. Le terme est souvent utilisé pour décrire l'ensemble de bases de stockage classifiées comme Big Data. Dans le monde du Big Data se trouvent aujourd'hui principalement Cassandra, MongoDB, NuoDB, Couchbase et VoltDB.

Références

1. "Big Data Use Cases". Amir Halfon. http://www.finextra.com/community/fullblog.aspx?blogid=6276