IBM BigInsights for Apache Hadoop

Gérer et explorer efficacement le Big Data pour exploiter tous les signaux

Points clés :

- Plateforme Hadoop prête à l'emploi pour le traitement, le stockage et l'analyse des données
- Analytique avancée de données structurées, semi-structurées et non structurées
- Outils de visualisation, de développement et d'administration pour une productivité optimale
- Accélérateurs d'application permettant une implémentation et un retour sur investissement plus rapides
- Intégration à des offres IBM éprouvées et également à d'autres solutions du marché

Apprivoiser le Big Data

IBM® Biglnsights for Apache™ Hadoop® permet aux entreprises d'exploiter de larges et complexes volumes de données, en relevant une multitude de défis métier. De manière générale, ces défis peuvent être classés en trois grandes catégories : efficacité opérationnelle, analytique avancée, exploration et découverte.

Performances opérationnelles

Pour gérer plus efficacement les performances et l'impact économique de l'augmentation du volume de données, il est possible d'utiliser conjointement des architectures intégrant différentes caractéristiques opérationnelles. Par exemple, de grandes quantités de données brutes d'un entrepôt de données peuvent être archivées dans un environnement analytique plutôt que dans un espace de stockage passif.

BigInsights permet d'améliorer les performances opérationnelles en modernisant (et non en remplaçant) l'environnement d'entrepôt de données. Celui-ci peut servir d'archive dotée d'une fonction de requête. Ainsi, les organisations peuvent stocker et analyser de grands volumes de données poly-structurées sans surcharger l'entrepôt de données. Centre de traitement préliminaire, également appelé « point de chute » de données, BigInsights permet aux entreprises d'explorer leurs données, de déterminer les éléments à forte valeur ajoutée et in fine d'extraire ces données. Le logiciel prend également en charge l'analyse ad hoc de grandes quantités de données pour l'exploration, la découverte et l'analyse.

Analytique avancée

Outre l'amélioration des performances opérationnelles, certaines entreprises souhaitent mener de nouvelles analyses, plus poussées, mais ne disposent pas des outils appropriés. Grâce à BigInsights, l'analytique n'est plus une étape distincte effectuée après le stockage de données. Au contraire, BigInsights, en association avec InfoSphere Streams, permet de réaliser des analyses en temps réel consistant à utiliser l'historique des modèles issus des données analysées au repos. BigInsights comprend également des fonctions d'analyse de texte avancée et des accélérateurs prédéfinis. Les entreprises peuvent donc utiliser ces fonctions analytiques pré-intégrées pour comprendre le contexte du texte de documents non structurés, effectuer des analyses de ressenti sur les données de réseaux sociaux ou obtenir des informations à partir de sources de données très variées.

Exploration et découverte

Les entreprises peuvent se sentir submergées par l'explosion du Big Data et rencontrer des difficultés à extraire des éléments d'informations essentielles. BigInsights contribue à instaurer un environnement parfaitement adapté à l'exploration et à la découverte des relations et corrélations entre les données. Il est alors possible d'obtenir de nouvelles perspectives et d'améliorer les résultats métier. Les Data Scientists peuvent analyser des données brutes issues de sources Big Data, ainsi que des données stockées dans les datawarehouses de la société et de plusieurs autres sources d'environnements de type bac à sable. Ensuite, ils sont en mesure d'associer les informations essentielles ainsi obtenues à d'autres données, en vue d'améliorer les informations et la prise de décisions des domaines opérationnel et stratégique.

Pour résumer : grâce à BigInsights, les entreprises peuvent enfin traiter des masses de données jusque là non exploitées et les explorer pour en extraire de précieuses informations, de manière optimale et évolutive.

Adopter Hadoop en entreprise

BigInsights for Hadoop est le fruit de l'association d'Apache Hadoop en open source et des innovations IBM. C'est un outil évolutif de traitement et d'analyse d'énormes quantités de données intégrant résilience et tolérance aux défaillances. IBM a élaboré des fonctions d'administration et de gestion simplifiées, des outils de développement complets et des fonctions analytiques performantes : ainsi, il est plus simple de se familiariser avec Hadoop.

L'un des plus gros défis lors de la conception d'applications à l'aide de solutions Hadoop en open source ou de tiers est le niveau élevé de compétence exigé. Avec BigInsights, ce n'est plus un problème : les deux grandes catégories de compétences liées au traitement de données, c'est-à-dire l'utilisation de tableurs et la programmation SQL, deviennent accessibles, car la création d'applications et l'extraction d'informations sont simplifiées.

Big SQL

Big SQL n'utilise pas Map-Reduce, mais un moteur SQL de traitement massivement parallèle (MPP) directement sur le cluster physique HDFS (Hadoop Distributed File System). Les performances et les fonctions d'exécution SQL sur Apache Hive 12 sont donc considérablement améliorées. Big SQL utilise un SQL standard pour permettre aux utilisateurs d'accéder au Big Data de la même manière qu'ils utilisent d'autres données relationnelles. BigInsights comprend également un tableau de bord intégré offrant une fonction d'interaction prête à l'emploi entre l'utilisateur et le Big Data. Il s'intègre de manière transparente via Big SQL à IBM Cognos® Business Intelligence pour une utilisation interactive de tableaux de bord et d'activités.

La puissance de Hadoop

BigInsights optimise Hadoop en open source grâce à la fonctionnalité d'entreprise et à l'intégration qui sont nécessaires pour répondre aux besoins métier essentiels. Les entreprises peuvent mener des travaux analytiques distribués à grande échelle sur des clusters de matériel serveur peu coûteux. Cette infrastructure utilise la plateforme Hadoop MapReduce pour traiter des ensembles de données très volumineux. Elle répartit les données sur de nombreux nœuds et coordonne leur traitement dans un environnement massivement parallèle**. Une fois les données stockées sur le cluster distribué, les systèmes peuvent exécuter efficacement les requêtes et l'analyse de données.**

Performance : selon les tests de performance, Big SQL exécute les requêtes **20 fois plus vite, en moyenne**, que Apache Hive 12. Les performances peuvent être multipliées par 70 pour les requêtes individuelles.

Compatibilité totale avec SQL : Big SQL 3.0 a exécuté avec succès l'ensemble des **99 requêtes TPC-DS** et **DES 22 requêtes TPC-H sans modification**. En revanche, Apache Hive 12 n'exécute que 43 des 99 requêtes TPC-DS sans modification.

L'accès aux lignes et colonnes Big SQL permet un contrôle d'accès aux lignes et colonnes, ou un « contrôle granulaire » cohérent avec la fonctionnalité offerte par un système RDBMS.

Accès fédéré aux données : Big SQL peut accéder aux données avec d'autres solutions que BigInsights. Cette fonction permet aux utilisateurs d'envoyer des requêtes distribuées à plusieurs sources de données à l'aide d'une seule instruction SQL.

Les administrateurs utilisent d'abord un outil d'installation par interface graphique, qui les invite à spécifier les composants facultatifs à installer et le mode de configuration de la plateforme. La progression de l'installation est présentée en temps réel et un système intégré de contrôle d'intégrité vérifie automatiquement le succès de l'installation. Ces fonctions d'installation avancées réduisent le temps nécessaire à l'installation et à l'optimisation. Les administrateurs peuvent ainsi se consacrer à d'autres projets essentiels.

Une fois que le cluster Hadoop est en place, ses fonctions de gestion des tâches permettent aux entreprises d'assurer le contrôle, des tâches BigInsights, des rôles utilisateur et de la surveillance de la sécurité et des indicateurs de performances clés (KPI). Le personnel technique peut facilement administrer la création, la soumission et l'annulation des tâches. Il peut également se tenir informé sur l'évolution de la charge de travail grâce à des tableaux de bord, des journaux et des moniteurs intégrés sur l'état des tâches. Ces outils fournissent des renseignements sur la configuration, les tâches, les tentatives et d'autres informations essentielles. De plus, BigInsights comprend des fonctions d'administration pour Hadoop Distributed File System (HDFS), IBM GPFS™ File Placement Optimizer (FPO), des tâches liées aux applications Big Data et à MapReduce, et la gestion en cluster.

Comme le montre la Figure 1, BigInsights for Hadoop comprend plusieurs fonctions d'entreprise. Les sections suivantes présentent de manière plus détaillée chaque domaine de ces fonctions.

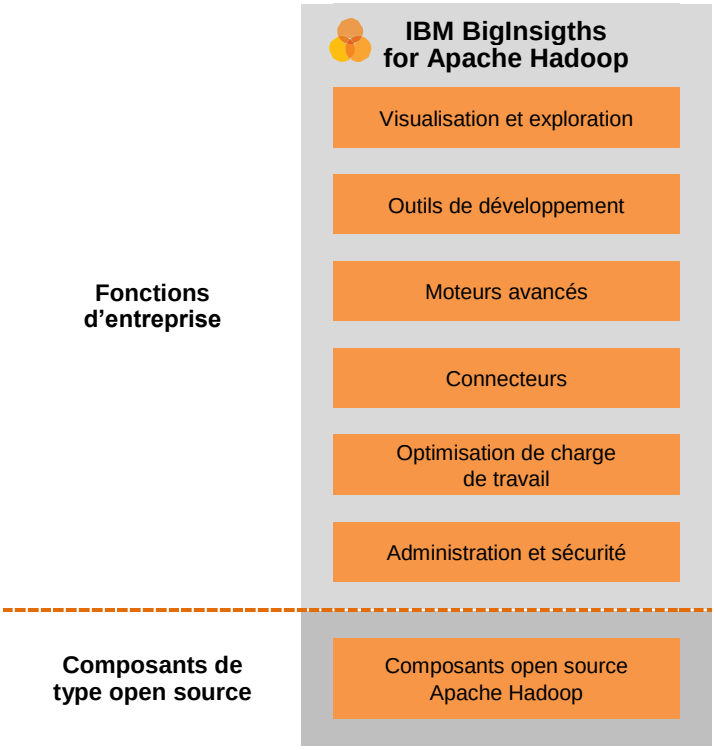

Figure 1. BigInsights ajoute des fonctions d'entreprise aux composants open source.

Essayez BigInsights gratuitement

BigInsights Quick Start Edition est une version gratuite hors production de BigInsights, disponible en téléchargement. Vous avez ainsi la possibilité d'explorer Hadoop sans restrictions de capacité ou de durée. Pour télécharger dès aujourd'hui votre Quick Start Edition, visitez le site : ibm.com/software/data/infosphere/biginsights/quick-start

Visualisation et exploration

BigInsights permet l'exploration et l'analyse ad hoc de toutes les données stockées sur la plateforme, ainsi que leur visualisation sous diverses formes.

BigSheets, tableaux de bord et exploration de données

BigSheets est un outil basé sur un navigateur qui se présente sous forme de tableur. Les Data Scientists comme les autres analystes métiers peuvent l'utiliser pour explorer, manipuler et analyser les Big Data.

BigSheets peut aider les professionnels à effectuer les tâches suivantes :

- Intégration et exploration de grandes quantités de données sous différents formats et structures.
- Extraction et enrichissement de données par l'analyse de texte.
- Exploration et visualisation de données à l'aide de graphiques et de tableaux croisés dynamiques.

BigInsights comprend également un tableau de bord centralisé permettant aux analystes métier d'extraire des informations de leurs données et de visualiser les résultats analytiques à grande échelle. Les administrateurs peuvent utiliser le tableau de bord pour surveiller les métriques de performance clés de leur cluster BigInsights for Hadoop.

Outils de développement

BigInsights utilise un environnement de développement familier, basé sur Eclipse, pour la création et le déploiement d'applications. Le logiciel fournit des éditeurs pour les composants Hadoop tels que Java™ MapReduce, Hive et Pig. Il comprend également une interface de programmation pour Big SQL, Oozie Workflows et Text Analytics.

BigInsights est également fourni avec des outils unifiés liés au cycle de vie de développement. Ainsi, les utilisateurs peuvent extraire des données d'Hadoop, les importer dans leur environnement de développement, puis développer, tester et déployer des applications sur le cluster.

Moteurs avancés et accélérateurs

BigInsights inclut un ensemble sophistiqué d'outils et de fonctions analytiques sans frais supplémentaires. Grâce à cette solution prête à l'emploi, les entreprises peuvent commencer rapidement à définir des modèles en s'appuyant sur leurs données et à créer des applications analytiques performantes et personnalisées. Celles-ci leur offriront des résultats et des perspectives parfaitement adaptés à leurs besoins métier.

Analyse de texte avancée

BigInsights inclut un puissant moteur d'analyse de texte développé par IBM Research. En utilisant une bibliothèque complète de règles ou en développant leurs règles personnalisées, les utilisateurs peuvent extraire et identifier rapidement les éléments d'intérêt dans les documents et les messages : noms de personnes, adresses e-mail, adresses postales, numéros de téléphone, adresses URL, fusions, alliances, entre autres.

Accélérateur d'analyse de données de réseaux sociaux

L'accélérateur d'analyse de données de réseaux sociaux permet aux utilisateurs d'analyser différents types de données de réseaux sociaux, en vue d'obtenir des informations clés à l'appui de la BI. Il est capable de capturer des informations cruciales sur les consommateurs, notamment le ressenti, l'intention d'achat et l'acquisition de produits/services. Il peut également obtenir des attributs démographiques : sexe, domicile, statut parental, statut marital, emploi, centres d'intérêt, sociétés dont la personne est cliente, produits en sa possession et intérêts liés aux produits.

Les sociétés peuvent exploiter ces attributs pour créer des applications servant plusieurs fonctions : génération de clients potentiels, fidélisation des clients/réduction de l'attrition, acquisition de clients et campagnes marketing ciblées.

Accélérateur d'analyse de données machine

L'accélérateur d'analyse de données machine peut intégrer, analyser et extraire divers types de données machine issues de sources telles que les fichiers journaux, les appareils intelligents et la télémétrie. Il permet de traiter ces données en quelques minutes, et non en plusieurs jours ou semaines. Les entreprises obtiennent ainsi des informations sur l'exploitation, les transactions et le fonctionnement du système. Ces résultats peuvent servir à dynamiser les performances opérationnelles de manière proactive, à résoudre ou identifier les principales causes des problèmes et à analyser les incidents. Ainsi, la société pourra éviter la dégradation ou l'interruption du service.

Connecteurs

Les technologies Big Data peuvent jouer un rôle important dans la chaîne logistique d'information de l'entreprise, mais uniquement si elles sont parfaitement intégrées aux systèmes existants. IBM l'a bien compris. C'est pourquoi BigInsights est doté de connecteurs ultra-rapides pour tous les types de données (structurées, non structurées et en temps réel) et toutes les sources (entrepôt de données, réseaux sociaux, données de journal,…). Les connecteurs d'intégration prédéfinis peuvent transférer les données vers des systèmes structurés et vers le système de fichiers Hadoop, alors que BigInsights peut importer directement des données non structurées.

BigInsights fournit des connecteurs au logiciel de base de données IBM DB2®, aux appliances analytiques IBM PureData™ Systems (technologie Netezza), à IBM InfoSphere Warehouse et à IBM Smart Analytics System. Ces connecteurs ultra-rapides permettent de simplifier et d'accélérer les tâches de manipulation de données. Les connecteurs standard JDBC (Java Database Connectivity) offrent la possibilité aux entreprises d'effectuer rapidement l'intégration à des systèmes de données et d'informations très variés, notamment Oracle, Microsoft® SQL Server, MySQL et Teradata.

De plus, IBM InfoSphere DataStage® inclut un connecteur permettant d'exploiter des données BigInsights dans le cadre d'une tâche InfoSphere DataStage ETL (Extract/Transform/Load) ou ELT (Extract/Load/Transform).

Optimisation de charge de travail

BigInsights comprend plusieurs fonctions contribuant à améliorer les performances, tout en optimisant leur adaptabilité et leur compatibilité au sein d'un environnement d'entreprise.

Planificateur d'affectation de 'workflow'

Les 'workflow' n'ont pas tous la même priorité. Le planificateur BigInsights offre un support d'affectation de 'workflow' adaptable aux tâches MapReduce, qui optimise le traitement en fonction d'une politique fixée par l'utilisateur. Le planificateur est une extension de Hadoop Fair Scheduler, conçu pour allouer à toutes les tâches, à long terme, une part équitable de ressources de cluster.

Technologie Adaptative MapReduce d'accélération de tâches

Les tâches s'exécutant sur Hadoop peuvent entraîner la création de multiples petites tâches qui consomment une quantité démesurée de ressources système. Pour y remédier, IBM a inventé une technique appelée Adaptive MapReduce. Conçue pour accélérer les petites tâches, elle modifie le mode de traitement des tâches MapReduce sans altérer leur création. Adaptive MapReduce s'intègre de manière transparente aux opérations MapReduce et aux opérations Hadoop liées à l'interface de programme d'application (API).

Administration et sécurité

Les exigences de l'entreprise en matière de sécurité doivent s'appliquer au Big Data, comme c'est le cas pour toutes les autres ressources d'information d'entreprise. BigInsights comprend plusieurs options sophistiquées permettant d'assurer la sécurité et la confidentialité des données.

Authentification

Plusieurs types d'authentification possibles s'offrent aux administrateurs : fichier plat, LDAP (Lightweight Directory Access Protocol) ou PAM (Pluggable Authentication Modules) pour la console Web BigInsights. Lors de l'authentification LDAP, le programme d'installation BigInsights communique avec une base de données d'identification permettant l'authentification. Les administrateurs peuvent ensuite octroyer l'accès à la console BigInsights en fonction de l'appartenance à des rôles, ce qui facilite la définition de droits d'accès pour des groupes d'utilisateurs.

Fonctions

BigInsights offre quatre niveaux de rôles utilisateur : administrateur système, administrateur de données, administrateur d'applications et utilisateur non administrateur. L'accès aux données et fonctions dépend du rôle affecté à chaque utilisateur.

Audit et sécurité

Il est possible d'exécuter les tâches MapReduce en utilisant des ID de comptes désignés, ce qui permet de renforcer la sécurité, le contrôle d'accès et l'audit. Par ailleurs, l'intégration de BigInsights au logiciel de sécurité de données IBM InfoSphere Guardium® permet aux entreprises de gérer les besoins liés à la sécurité et à l'audit de Hadoop de la même manière qu'elles le font pour les sources de données structurées traditionnelles.

BigInsights est également compatible avec le protocole d'authentification de service à service Kerberos : la sécurité est renforcée pour prévenir les attaques de type « middle man ».

Amélioration de l'intégration d'entreprise

IBM Watson Explorer

BigInsights inclut une licence restreinte de Watson Explorer, qui permet aux entreprises de découvrir, parcourir et visualiser de grands volumes d'informations structurées et non structurées sur les systèmes et les référentiels de données d'entreprise. Ce logiciel fournit également un point d'accès économique et efficace permettant d'explorer les atouts des technologies Big Data par le biais d'une puissante plateforme de développement d'applications utilisant les données d'entreprise existantes.

InfoSphere Streams

BigInsights inclut une licence restreinte d'InfoSphere Streams, qui permet d'effectuer des analyses de données continues en temps réel. InfoSphere Streams est un système de traitement des flux d'entreprise capable d'extraire des informations exploitables de données en mouvement, tout en assurant leur transformation et leur transfert vers BigInsights à grande vitesse. Cela permet aux entreprises de capturer et d'exploiter les données métier en temps réel, par l'intégration, l'analyse et la mise en corrélation d'informations dès leur arrivée. Les performances de traitement en sont nettement améliorées.

Cognos Business Intelligence

BigInsights inclut une licence restreinte de Cognos Business Intelligence, qui permet aux utilisateurs d'entreprise d'accéder aux informations dont ils ont besoin et de les analyser. Ainsi, ils peuvent améliorer la prise de décisions, obtenir des informations plus pertinentes et mieux gérer les performances. Cognos Business Intelligence inclut des logiciels de requête, de reporting, d'analyse et des tableaux de bord, ainsi que des logiciels permettant de collecter et d'organiser des informations issues de plusieurs sources.

InfoSphere Master Data Management

Pour les utilisateurs effectuant des analyses client, BigInsights utilise le moteur de mise en correspondance probabiliste d'InfoSphere Master Data Management pour mettre en correspondance et relier des informations client directement dans Hadoop, à grande vitesse. Un identifiant propre à chaque client garantit une meilleure précision de l'analyse d'informations.

Conclusion

BigInsights for Hadoop repose à 100 % sur Apache Hadoop Open Source et comprend des fonctions professionnelles permettant de prendre en charge tous les cas d'utilisation de Big Data. IBM optimise l'expérience Hadoop en apportant la haute disponibilité, la formation, le support et les services nécessaires à la réussite du déploiement et à un retour sur investissement rapide.

Pour plus d'informations

Pour en savoir plus sur IBM BigInsights for Apache Hadoop, contactez votre représentant ou partenaire commercial IBM ou visitez le site Web suivant :
ibm.com/software/data/infosphere/biginsights

Compagnie IBM France
17 avenue de l'Europe
92275 Bois Colombes Cedex

Imprimé en France
Mars 2015

Recyclable

IMD14385-FRFR-04