

Informatica HParser

Simplifier le développement dans Hadoop

Arun C. Murthy

Co-fondateur

Projet MapReduce dans Apache

Hadoop

Hortonworks

« Alors que nous faisons progresser l'infrastructure Apache Hadoop pour le stockage, le traitement et l'analyse des grands volumes de données, l'approche d'Informatica HParser permettant l'analyse parallèle des clusters Hadoop disponibles dans MapReduce constitue un apport unique à la communauté Apache Hadoop. Dans la distribution Hortonworks, Informatica HParser utilise l'infrastructure de programmation Apache Hadoop de manière unique et offre une évolutivité linéaire. »

BÉNÉFICES

- Gérer facilement plusieurs normes industrielles, documents binaires et données hiérarchisées complexes
- Développer facilement à l'aide d'une interface utilisateur intuitive permettant d'afficher les échantillons de données dans leur format d'origine et sous forme de fichier texte
- Analyser n'importe quel format de données depuis Hadoop, à l'aide d'un moteur de transformation unique accessible par un simple appel pour le développeur MapReduce
- Transformer rapidement des formats structurés et semi-structurés à l'aide d'un environnement de développement graphique

Gestionnaire de données universel pour Hadoop

Informatica HParser offre un accès aux données et aux formats de fichiers les plus complexes de Hadoop, réduisant ainsi de 70 % la durée et les coûts de développement. À l'aide d'Informatica Data Transformation Studio, un environnement d'analyse graphique basé sur des exemples, l'utilisateur peut facilement définir la méthode de conversion des données complexes en formats plats, pratiques à utiliser et exploitables par traitement Hive, PIG ou MapReduce. Simplifiant grandement le développement de la logique Hadoop, Informatica HParser repose sur un moteur accessible par un simple appel pour analyser et gérer n'importe quel format, permettant ainsi au développeur de travailler sur une seule couche d'abstraction de données.

Grâce à Informatica HParser, les entreprises peuvent :

- Définir des règles de parsing complexes dans un environnement graphique basé sur des exemples, avec prise en charge immédiate des logs, prise en charge intégrée des formats binaires et transformations packagées pour les formats industriels
- Déployer et réutiliser immédiatement les transformations dans Hadoop MapReduce
- Normaliser plusieurs formats de données dans une abstraction de données unique, à l'aide d'une transformation unique

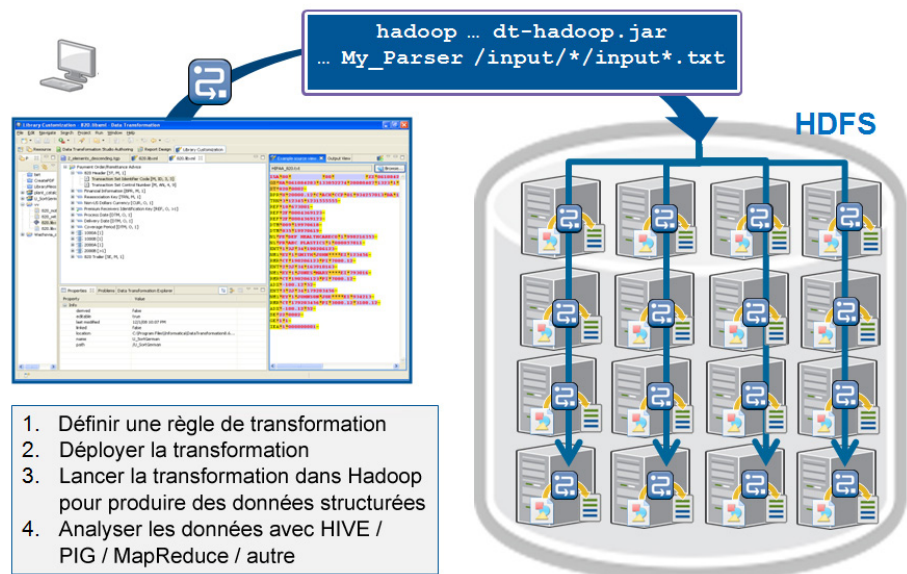


Figure 1. Analyse et gestion des données par HParser dans Hadoop

Principales fonctionnalités

Prise en charge d'un grand nombre de formats de données

Prise en charge de plusieurs normes industrielles. Le traitement de normes et de formats de données spécifiques de l'industrie – tels que le format EDI pour le secteur industriel, SWIFT, NACHA et SEPA pour les paiements, ACORD pour le secteur des assurances, ASN.1 pour les opérateurs de télécommunications, HL7 pour le secteur de la santé, etc. – présente une difficulté particulière. Ces normes, généralement instaurées par les groupes industriels et les organisations publiques, sont en évolution constante et introduisent des formats nouveaux et améliorés. La plupart de ces normes font l'objet d'au moins une nouvelle version par an, ce qui implique que toute grande initiative d'analyse de grands volumes de données sur plusieurs années doit prendre en charge plusieurs versions et variations. Grâce à notre grand choix de bibliothèques, de versions et de messages, Informatica HParser élimine ce problème en fournissant des mises à jour régulières des normes nouvelles et existantes, dès leur émission par les entreprises et les organisations publiques. Avec ces mises à jour, le processus actuel est en mesure de prendre en charge les nouveaux formats dès qu'ils sont disponibles.

Prise en charge unique des documents binaires : Word, Excel, PDF. Les organisations stockent des quantités considérables de données dans des documents (dossiers juridiques et contrats sous forme de documents Word et PDF, par exemple) et des rapports financiers et des prévisions dans Excel. Informatica HParser offre une prise en charge immédiate de ces documents binaires. Les utilisateurs peuvent ainsi en traiter et en extraire les données pertinentes pour les importer dans Hadoop.

Prise en charge complète des données hiérarchiques. Certains formats tels que XML et JSON augmentent la complexité des données hiérarchiques. Il est indispensable de pouvoir traiter efficacement les données de hiérarchie profonde et prendre en charge les schémas et les structures avancés afin de gérer de manière optimale les données complexes dans ces formats. Informatica HParser offre une prise en charge native des formats XML et JSON ainsi qu'une approche optimisée pour l'extraction de données de structures hiérarchisées.

Prise en charge des logs. À l'aide d'un moteur breveté de transformation basée sur des spécifications, Informatica HParser simplifie la définition de spécifications des logs, y compris de journaux hiérarchiques, délimités et positionnels. Ces spécifications peuvent également être exploitées pour analyser et extraire les données des logs, dont les logs Web, d'enregistrements détaillés des appels, mainframe et propriétaires.

Développement aisé grâce à l'interface utilisateur intuitive

La fonctionnalité unique de transformation basée sur des exemples d'Informatica HParser augmente considérablement la productivité par rapport aux approches de traitement et d'analyse de données plus traditionnelles, telles que les applications Java, Perl et XSLT. Lorsque la source est un fichier PDF, une feuille de calcul Excel, un modèle COBOL ou un message Bloomberg, SWIFT ou ASN.1, l'interface utilisateur HParser permet l'affichage d'un échantillon de données dans son format d'origine et sous forme de fichier texte. L'approche basée sur des exemples favorise la mise en œuvre de règles de parsing, et fournit un retour immédiat, sans qu'aucune compilation ni aucun déploiement ne soit nécessaire.

Intégration et déploiement dans MapReduce

Informatica Data Transformation permet un déploiement universel dans différents environnements et une intégration transparente dans toute une palette d'infrastructures logicielles, dans toutes les topologies. Cette solution se connecte directement à n'importe quelle plate-forme SOA, EAI, B2B ou d'intégration de données. Informatica HParser améliore la flexibilité de déploiement en s'intégrant parfaitement à Hadoop MapReduce. Le développeur MapReduce peut accéder au moteur HParser par une simple commande, afin d'analyser n'importe quel format de données dans Hadoop.

Abstraction de données rapide pour le développeur Hadoop (MapReduce)

Les scénarios d'analyse avancée de grands volumes de données dépendent de la capacité à traiter les données issues de plusieurs sources. Informatica HParser fournit un environnement de développement graphique pour convertir rapidement ces formats structurés et semi-structurés en un format pivot utilisable et plat. L'utilisation d'HParser comme moteur de transformation unique, à la place de plusieurs gestionnaires de données codées manuellement, permet au développeur MapReduce de développer un programme unique qui s'adapte facilement aux différentes variations de données.

Pour en savoir plus

Pour en savoir plus sur la plate-forme Informatica, visitez le site www.informatica.com/fr ou contactez Informatica au 01 42 04 89 00.

À propos d'Informatica

Informatica Corporation (NASDAQ : INFA) est le leader des fournisseurs indépendants de solutions d'intégration de données. Les sociétés du monde entier ont choisi Informatica pour gagner un avantage concurrentiel certain grâce à des données pertinentes et fiables répondant en temps voulu à leurs principaux impératifs métiers. Plus de 4 500 entreprises dans le monde s'appuient sur Informatica pour les solutions d'intégration des données, de qualité des données et de gestion de grands volumes de données, permettant d'accéder aux informations hébergées sur site et dans le cloud, de les intégrer et de renforcer leur fiabilité. Pour en savoir plus, appelez le 01 42 04 89 00 ou visitez notre site www.informatica.com/fr. Entrez en contact avec Informatica via les sites <http://www.facebook.com/InformaticaCorporation>, <http://www.linkedin.com/company/informatica> et <http://twitter.com/InformaticaFr> ou Viadeo : <http://bit.ly/zj8fzZ>.

INFORMATICA[®]
The Data Integration Company™

Siège mondial, 100 Cardinal Way, Redwood City, CA 94063, États-Unis
France : +33 1 42 04 89 00

© 2011 Informatica Corporation. Tous droits réservés. Imprimé aux États-Unis. Informatica, le logo Informatica et The Data Integration Company sont des marques commerciales ou déposées appartenant à Informatica Corporation aux États-Unis et dans d'autres pays. Tous les autres noms de sociétés et de produits sont la propriété de leurs détenteurs respectifs et peuvent avoir fait l'objet d'un dépôt de marque. Première publication : novembre 2011

1855FR (01/11/2011)