



Talend Open Profiler

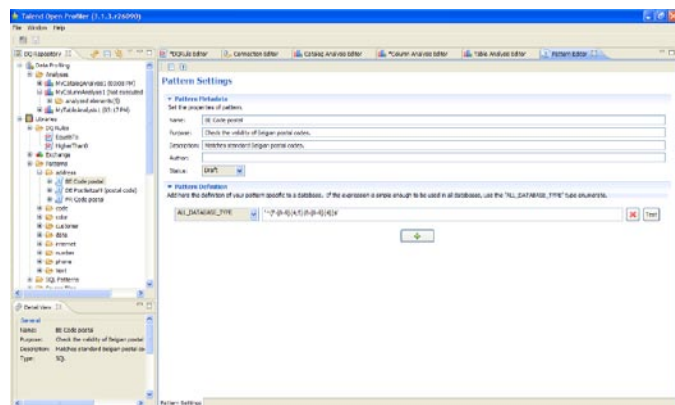
Premier outil open source de profiling de données

Le Data Profiling se définit par l'analyse des données présentes dans différentes sources de données et la production de statistiques et d'informations sur ces données. Le Data Profiling permet d'évaluer le niveau de qualité des données contenues dans le système d'information selon un ensemble de métriques et de seuils définis au préalable.

Un outil robuste

Talend Open Profiler est un outil sophistiqué mais simple d'utilisation, qui permet de profiler le contenu, la structure et la qualité des structures de données les plus complexes. Les utilisateurs métier ou les équipes en charge de la gestion des données peuvent ainsi effectuer toutes sortes d'analyses à l'aide d'un ensemble d'indicateurs, de motifs et de règles pour chaque élément de données à analyser ou superviser. Talend Open Profiler permet d'analyser les données en continu et d'analyser l'évolution dans le temps des données sources afin d'améliorer la qualité des données.

Ces indicateurs peuvent être des indicateurs statistiques simples ou avancés, des indicateurs alphanumériques ou numériques, mais aussi des données de synthèse ou une distribution statistique des données. Les motifs sont des expressions prédéfinies ou personnalisées, qui déterminent la forme attendue des données analysées tandis que les règles métier de qualité de données permettent de définir des seuils personnalisés et des plages de valeurs. Ces indicateurs, motifs et règles ont pour but de mettre en exergue toutes les valeurs anormales.



Talend Open Profiler produit des rapports et des graphiques sophistiqués qui permettent aux utilisateurs d'évaluer rapidement le niveau de qualité des données et de voir le statut des indicateurs définis. De plus, un explorateur de données embarqué permet aux utilisateurs de sonder les bases de données analysées.

Fonctionnalités clés

Découverte des métadonnées

Talend Open Profiler se connecte aux bases de données pour inspecter leurs structures et centralise la description de leurs données dans son Metadata Manager. Les utilisateurs disposent d'un système de filtrage avancé pour limiter l'analyse à une partie des tables ou colonnes, afin d'optimiser les temps de connexion en cas de grand nombre de tables et d'analyser plus efficacement les données les plus pertinentes.

Les métadonnées sont alors utilisées par les analystes pour réaliser des comparaisons et des analyses de bases de données et pour établir des métriques et indicateurs. Ces derniers pourront ensuite servir aux utilisateurs pour évaluer leur besoin en termes de qualité de données et prendre des mesures adéquates de nettoyage de données, d'intégration de données ou de gouvernance de données.

Column Name	Data Type	Type Name	Column Size	Decimal Digits	Radix	Is Null Allowed	Remarks	Defa
TEST_CHAR	1	char	10	0	10	0		
TEST_CHAR_NULL	1	char	10	0	10	1	<null	
TEST_DATE	91	date	0	0	10	0		
TEST_DATE_NULL	91	date	0	0	10	1	<null	
TEST_DOUBLE	8	double	22	0	10	0		
TEST_DOUBLE_NL	8	double	22	0	10	1	<null	

Columns for "TEST_DATAPROFILER"."TEST_COUNT"

Règles métiers personnalisées

Les règles métiers sont des critères spécifiques, des seuils ou des plages de valeurs qui permettent d'identifier les enregistrements concordants, illogiques (par exemple : âge saisi < 0 ou décimal) ou ne correspondant pas aux valeurs attendues.

Un assistant dédié permet de définir et personnaliser facilement des règles métiers de qualité de données à l'aide du standard SQL et permet, en utilisation avancée, de créer des jointures pour les besoins plus complexes. Ces règles sont utilisées pour définir des seuils de valeur attendue pour les indicateurs de qualité de données. La plage de valeurs ou valeur définie sert à mesurer la qualité des données de la table sélectionnée.

Motifs

Les motifs sont des données de référence contre lesquelles les données analysées sont comparées. Une bibliothèque de motifs prédéfinis est disponible nativement pour résoudre les problèmes de qualité les plus fréquents.

En outre, des motifs entièrement personnalisés peuvent être créés à partir d'expressions régulières et de requêtes SQL pour une analyse des données plus précise.

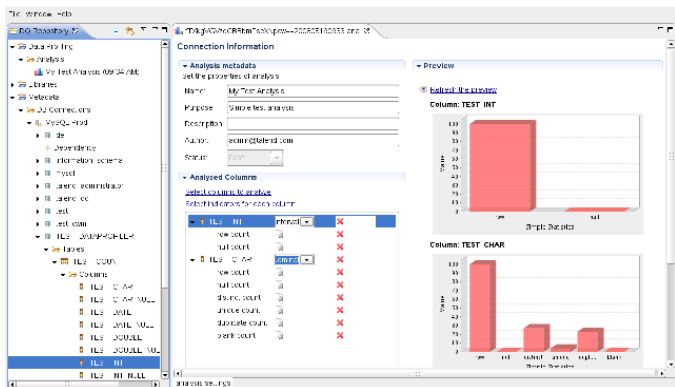
Les utilisateurs peuvent partager leurs motifs et récupérer les motifs développés par les autres utilisateurs de la communauté Talend via la plateforme Talend Exchange accessible directement depuis le studio de Talend Open Profiler. Les expressions régulières ou requêtes SQL peuvent aussi être importées à partir d'un fichier CSV lorsque le nombre de motifs est important.

Indicateurs

Les indicateurs sont le résultat de la mise en place des différents motifs. Ils définissent le contenu, la structure et la qualité des données analysées et peuvent résulter d'opérations simples comme complexes de mise en correspondance des données notamment.

Un ensemble d'indicateurs est fourni nativement dans le Studio afin d'aider les utilisateurs à prendre en main Talend Open Profiler. Ils incluent :

- **Simple statistics** : analyse statistique sur le nombre d'enregistrements tel que le nombre de lignes, de valeurs nulles, de valeurs uniques et distinctes, de doublons et de champs vides.
- **Text statistics** : analyse des caractéristiques des champs texte, telles que les longueurs minimale, maximale et moyenne.
- **Summary statistics** : analyse statistique sur des données numériques, dont le calcul de la médiane, de la moyenne, du premier et dernier quartile et des intervalles.
- **Advanced statistics** : recherche des valeurs les plus probables et les plus fréquentes et création de tableaux de fréquence basés sur ces valeurs.
- **Pattern frequency** : calcul de la fréquence des enregistrements pour chaque modèle (plus fréquents et moins fréquents).
- **Soundex frequency** : indexation des enregistrements basés sur la phonétique et la prononciation.



Des assistants dédiés permettent aux utilisateurs de définir leurs propres indicateurs à l'aide de requêtes SQL ou Java afin de mettre en place de nouvelles métriques de qualité de données ou de répondre à des besoins spécifiques.

Interprétation

Pour chacun des éléments (table, colonne, donnée ou indicateur) sélectionné, Talend Open Profiler produit des rapports et des graphiques sophistiqués qui permet à l'utilisateur d'évaluer rapidement les résultats de leurs profiling directement dans l'éditeur d'analyse du Studio.

Avantages

En réalisant une analyse performante des données contenues dans les bases de données, Talend Open Profiler permet aux utilisateurs de juger du niveau de qualité de leurs données

d'obtenir des informations descriptives sur ces données. Connaître ses données est essentiel pour chaque projet d'intégration de données, et plus généralement pour chaque projet de données (développement applicatif, interface Web 2.0, etc.). C'est en comprenant clairement la structure, la nature et la qualité de ces données et en prenant en compte les spécificités locales ainsi que les exceptions métier mises en avant par le data profiling, que les analystes peuvent être en mesure de mener à bien leurs projets de gestion de données.

De plus, Talend Open Profiler est l'outil idéal pour la recherche de données erronées dans tous les systèmes. Il fournit une distribution statistique ou par domaine des enregistrements, facilitant considérablement la recherche d'éléments spécifiques.

Talend Open Profiler est doté d'une interface graphique et ne requiert aucune connaissance technique des bases de données ou de la structure des fichiers, car il fournit un accès non technique aux données. Les analystes fonctionnels ou autres responsables non techniques peuvent donc accéder aux informations critiques relatives aux données. En suivant régulièrement les métriques, ils peuvent ainsi observer l'évolution de la qualité des données.

Data profiling et intégration de données

Le profiling des données est une discipline spécifique qui permet aux utilisateurs d'évaluer la qualité de leurs données, d'une part, et de connaître la nature de ces données, d'autre part. Une bonne compréhension du niveau de qualité des données traitées ainsi que des zones à risque est une des clés du succès des projets d'intégration de données. En outre, les informations collectées avec Talend Open Profiler sont déterminantes pour la conception même des processus d'intégration de données. En effet tout processus d'intégration de données inclut généralement des opérations de qualité des données : validation, filtrage et déduplication des données, traitement des rejets, notification d'erreurs, etc. La solution de qualité de données d'entreprise, Talend Data Quality, fournit tous les outils nécessaires pour ces opérations de redressement et de normalisation des données.

Une communauté active d'utilisateurs

Les utilisateurs de Talend Open Profiler peuvent trouver un soutien actif de la part des nombreux membres de la communauté via les outils en ligne sur TalendForge :

- Forum: <http://www.talendforge.org/forum/>, le site où toutes les questions trouvent une réponse.
- Wiki: <http://www.talendforge.org/wiki/>, fournit toutes les informations utiles pour l'installation et l'utilisation de Talend Open Profiler.
- Bugtracker: <http://www.talendforge.org/bugs/>, l'outil de report des problèmes rencontrés dans le logiciel.
- Exchange: <http://talendforge.org/exchange/>, lieu de partage des contributions et développements externes.
- Babili: <http://www.talendforge.org/babili/>, portail de localisation et de traduction de tous les produits Talend.

Grâce à la communauté, Talend Open Profiler est désormais disponible en plus de 15 langues.