OFFRE DE SERVICE DATA QUALITY TOOLS

Data Quality Tools Service

Depuis fin 2009, Smals dispose d'outils de Data Quality. En 2008, le cahier des charges a été publié en deux phases puis testé de manière extensive. La solution qui a été choisie est le système Trillium Software (voir également Gartner Magic Quadrants pour les outils de Data Quality des dernières années).

Ces outils nous permettent de mener des projets, des analyses, des migrations et des intégrations de données, faisant face aux problématiques mentionnées plus bas (voir : Applications).

A cette fin, un moyen de production (DQRS) a été créé.

Applications

Plusieurs DQRS (le nombre est en fonction de la taille et de la complexité des bases de données à analyser) permettent d'effectuer un data profiling approfondi, une détection de doubles ou des incohérences, un parsing et un cleansing de l'information du nom et de l'adresse ou de toute autre information string, même la validation de l'adresse.

- **Data profiling**

  Tous les modèles et valeurs, dépendances et clés présents sont répertoriés. Ceci se fait en examinant les données et la structure de données et en les confrontant à la documentation et aux métadonnées disponibles.

  L'objectif est ainsi de trouver des données non conformes et de compléter et de corriger la documentation ou les métadonnées défaillantes, incomplètes ou obsolètes. De cette façon, on remarque le manque de standardisation ainsi que les cas où les règles business rules et les data rules sont violées.

  Le résultat obtenu avec des outils de Data Quality est exploitable et supporte la recherche de solutions, la proposition de standards et la concession de standards et de business rules.

- **Détection de doubles**

  Des doubles présents dans des bases de données importantes sont détectés de manière efficace et flexible puis catégorisés en typologies. Ceci permet de se concerter beaucoup plus facilement avec le business sur ce qui peut ou doit être un double, et ce qui peut être autorisé ou non dans les bases de données.

  Il est en outre possible de marquer des modèles match suspects en vue d'un traitement spécial (manuel), par exemple dans le cadre de la détection de fraude.

- **Détection d'incohérences**

  Ceci se fait entre deux sources de données ou plus.

  Tout d'abord, on effectue un profiling des différentes sources, ce qui permet de répertorier les différences de structure, de standardisation et de business rules. Ensuite, on effectue un fuzzy matching entre les deux sources.

  Ceci permet d'une part de détecter les doubles dans chaque source, d'autre part de trouver les liens (manquants) entre deux sources, même lorsque aucune clé n'est présente. Ceci a son importance dans le cas de contrôles d'exhaustivité, de détection de fraude, ...

- **Adresvalidatie, naam- en adres-cleansing**

- **Validation de l'adresse, du nom et cleansing de l'adresse**

  L'information sur le nom et l'adresse peut être parsée et standardisée dans un projet de qualité. Par la suite, il est possible de faire un match par rapport à un fichier de référence mis à jour plusieurs fois par an et qui est considéré comme une source authentique. Ainsi, on peut même passer à la validation de l'adresse. Après cette étape d'enrichissement, nous pouvons effectuer une détection de doubles et d'incohérences.

  Un tel projet de qualité est paramétré dans un GUI et peut être planifié comme projet batch pour être exécuté. En fonction des moyens à disposition par projet, il est aussi possible d'obtenir une intégration en ligne.

  Pour l'instant, la région **belge** est supportée dans les trois langues nationales, mais ceci peut éventuellement être étendu avec les bases de connaissances de toutes les régions mondiales qui seraient opportunes en fonction des moyens disponibles par projet (prix de revient supplémentaire non compris dans le prix d'un DQRS).

Des **combinaisons et** des **variantes** sont possibles sur ces quatre options et évidemment, des fonctionnalités peuvent être appliquées sur **toute sorte d'information**, non pas uniquement sur l'information du nom et de l'adresse.

Avantages

- **Accélérer les phases d'analyse**

  Très rapidement, et dans n'importe quel contexte, des problèmes de qualité des données peuvent être constatés. Par exemple: le manque de standardisation, la présence de doubles, la violation de business rules,…

  Ceci permet de mieux estimer ce qui doit être fait et l'effort que ceci nécessitera.

  L'analyse est suivie d'un moment de concertation avec le business afin de définir une stratégie de solution. Il est alors examiné si les problèmes doivent être traités avec des outils de Data Quality, de manière partiellement automatique ou non.

- **Itérer mieux et plus rapidement avec des connaisseurs business**
- **Fournir de meilleurs développements, opérer un coût de maintenance plus bas**

  L'on fournit de meilleurs développements, qui tiennent compte des problèmes de qualité des données et qui évitent en production des problèmes autrement imprévus pendant le développement ou l'analyse.

  Ceci se fait en faisant valider la stratégie, les méthodes et les résultats et en répondant aux change requests.

- **Estimer plus précisément les risques et l'effort requis**
- **Meilleure préparation des migrations de données, mieux palier aux difficultés liées a l'intégration de données**

Voorbeelden

Source	match type	Denomination	Adres	Boite	Postcd	Commune	Cdpays
L	100	PROJEKTSERWIS WANDA LITTY	BOHATEROW MODLINA 63	43	05-100	NOWY DWOR MAZ	122
L	100	PROJEKTSERWIS WANDA LITTY	BOHATEROW MODLINA 63	43	05-100	NOWY DWOR MAZ	122
L	100	PROJEKTSERWIS WANDA LITTY	BOHATEROW MODLINA 63	43	05-100	NOWY DWOR MAZ	122
L	100	PROJEKTSERWIS WANDA LITTY	BOHATEROW MODLINA 63	43	05-100	NOWY DWOR MAZ	122
R	115	PROJEKT SERWIS (LUTY WANDA)	UL. BOHATEROW MODLINA 63	42	05-100	NOWY DROW MZAOWIE	PL
R	115	PROJEKT SERWIS LUTY WANDA NOWY DWOR	UL BOHATEROW MODLINA 63	43	05-100	NOWY DWOR MAZOWIE	PL
R	135	PROJEKT SERWIS LUTY WANDA	BOHATEROW MODLINA 63/43		05-100	NOWY DWOR MAZ	PL
R	135	PROJEKT SERWIS LUTY WANDA	BOHATEROW MODLINA 63/43		05-100	NOWY DWOR MAZ	PL
R	106	PROJEKT SERWIS WANDA LUTY	BOHATREROW 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	106	PROJEKT SERWIS WANDA LUTY	BOHATREROW MODLINA 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	128	PROJEKT SERWIS WANDA LUTY	BOHATREROW MODLINA 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	128	PROJEKT SERWIS WANDA LUTY	BOHATREROW MODLINA 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	128	PROJEKT SERWIS WANDA LUTY	BOHATREROW MODLINA 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	128	PROJEKT SERWIS WANDA LUTY	BOHATREROW MODLINA 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	128	PROJEKT SERWIS WANDA LUTY	BOHATREROW MODLINA 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	128	PROJEKT SERWIS WANDA LUTY	BOHATREROW MODLINA 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	128	PROJEKT SERWIS WANDA LUTY	BOHATREROW MODLINA 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	128	PROJEKT SERWIS WANDA LUTY	BOHATREROW MODLINA 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	128	PROJEKT SERWIS WANDA LUTY	BOHATREROW MODLINA 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	128	PROJEKT SERWIS WANDA LUTY	BOHATREROW MODLINA 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	128	PROJEKT SERWIS WANDA LUTY	BOHATREROW MODLINA 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	128	PROJEKT SERWIS WANDA LUTY	BOHATREROW MODLINA 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	128	PROJEKT SERWIS WANDA LUTY	BOHATREROW MODLINA 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	128	PROJEKT SERWIS WANDA LUTY	BOHATREROW MODLINA 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	138	SOCIETE PROJEKTSERWIS	BOHATERROW MODLINA 63/43	N/A	05-100	NOWY DWOR MAZOWIE	PL
R	138	SOCIETE PROJEKTSERWIS	BOHATERROW MODLINA 63/43	N/A	05-100	NOWY DWOR MAZOWIE	PL

Des données réelles, le fuzzy matching et la détection d'incohérences entre 2 sources (L et R) – un groupe de doubles est affiché ainsi que le lien qui peut être établi, sans clé, entre les 2 registres.

C Postcode	Tq Gout Postal Code	Straatnaam Voll	Tq Gout Street Name	Huisnummer	Pr House N...	Gemeentenaam	Tq Gout Postal City
1020	1020	RUE E VANDER AA	RUE ERNEST VANDER AA	1	1	Brussel	BRUSSEL
1020	1020	rue Vander Aa	RUE ERNEST VANDER AA	3	3	Bruxelles	BRUXELLES
1050	1050	91 R VAN AA	RUE VAN AA	_	91	Elsene	ELSENE
1050	1050	27 R.VAN AA	RUE VAN AA	_	27	Elsene	ELSENE
1050	1050	RUE VAN AA	RUE VAN AA	94	94	Ixelles	IXELLES
1050	1050	RUE VAN AA	RUE VAN AA	94	94	Elsene	ELSENE
1020	1050	rue Van Aa	RUE VAN AA	2	2	Bruxelles	IXELLES
1050	1050	2 R VAN AA	RUE VAN AA		2	Ixelles	BRUXELLES
1000	1000	R JOSEPH II 40	RUE JOSEPH II		40	Bruxelles	BRUXELLES
1000	1000	rue Joseph II 71 (...	RUE JOSEPH II		71	Bruxelles	BRUXELLES
1040	1000	Rue Joseph II	RUE JOSEPH II	71	71	Brussel	BRUSSEL
1040	1000	Rue Joseph II 5-7	RUE JOSEPH II		5-7	Bruxelles	BRUXELLES
1040	1000	Rue Joseph II 67A	RUE JOSEPH II		67A	Bruxelles	BRUXELLES
1030	1000	rue JOSEPH II, 114 -	RUE JOSEPH II	116	114 - 116	Schaarbeek	BRUXELLES

Des données réelles, le cleansing (standardisation et matching) d'adresses – le code postal est corrigé, le nom de rue est standardisé, les différents éléments d'adresses sont classés correctement (parsing), le nom de commune est corrigé, les doublons sont détéctés et organisés en clusters.