DIENSTENAANBOD DATA QUALITY TOOLS SERVICE

Data Quality Tools Service

Smals beschikt sinds eind 2009 over Data Quality Tools. Het lastenboek met twee fasen werd in 2008 gepubliceerd en breedvoerig getest. De verkozen oplossing is het Trillium Software System (zie ook de Gartner Magic Quadrants voor Data Quality Tools van de laatste jaren).

Met behulp van deze tools kunnen we analyses, projecten, data migraties en data integraties ondersteunen waarin onderstaande problematieken centraal staan (zie Toepassingen).

Hiertoe werd een productiemiddel (DQRS) gecreëerd.

Toepassingen

In een aantal DQRS'en (in functie van de grootte en complexiteit van de te analyseren databanken) kan men een diepgaande data profiling uitvoeren, een dubbeldetectie of een incoherentiedetectie, een parsing en cleansing van naam- en adresinformatie of van om het even welke andere string-informatie, zelfs adresvalidatie.

- **Data profiling**

  Men brengt alle aanwezige waarden en patronen, dependencies en keys in kaart. Dit gebeurt door de gegevens en gegevensstructuur door te lichten en te toetsen aan de beschikbare documentatie en metadata.

  Het doel is zo niet-conforme data te vinden en gebrekkige, onvolledige of niet meer bijgewerkte documentatie/metadata te vervolledigen en accuraat te maken. Hieruit komt het gebrek aan standaardisatie en het overtreden van business rules en data rules naar voren.

  Het resultaat behaald met data quality tools is exploiteerbaar en ondersteunt het zoeken naar oplossingen, het voorstellen van standaarden en het afdwingen van standaarden of business rules.

- **Dubbeldetectie**

  Men ontdekt op efficiënte en flexibele wijze dubbels in grootschalige databanken en categoriseert deze in typologieën. Dit laat toe om veel gemakkelijker met de business te overleggen wat al dan niet een dubbel moet of mag zijn, en wat al dan niet toegelaten mag worden in de databanken.

  Ook is er de mogelijkheid om verdachte matchpatronen te laten vlaggen voor speciale (manuele) behandeling, bv. in het kader van fraudedetectie.

- **Incoherentiedetectie**

  Dit vindt plaats tussen 2 of meer databronnen.

  Eerst voert men een profiling uit van de verschillende bronnen die de verschillen in structuur, standaardisatie en business rules in kaart kan brengen. Vervolgens voert men een fuzzy matching uit tussen de twee bronnen.

  Dit laat toe om enerzijds dubbels te detecteren in elk van de bronnen, anderzijds om (ontbrekende) links tussen 2 bronnen te vinden, zelfs indien geen sleutel aanwezig is. Dit is van belang bij volledigheidscontroles, fraudedetectie, ...

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

  Naam- en adres informatie kan in een kwaliteitsproject geparsed en gestandaardiseerd worden. Vervolgens kan worden gematched ten opzichte van een meermaals per jaar bijgewerkt referentiebestand dat als authentieke bron wordt beschouwd. Zo kan men zelfs tot adresvalidatie overgaan. Na deze verrijkingsstap kan men een dubbel- incoherentiedetectie uitvoeren.

  Zulk kwaliteitsproject wordt in een GUI geparametriseerd en kan als batch-project ter uitvoering gescheduled worden. In functie van beschikbare middelen per project kan men ook een online integratie bekomen.

  Voorlopig is de **Belgische** regio ondersteund in de drie landstalen, maar dit kan eventueel uitgebreid worden met de knowledge bases van alle wenselijke regio's wereldwijd, in functie van de beschikbare middelen per project (bijkomende kostprijs niet inbegrepen in de prijs van een DQRS).

Op deze vier mogelijkheden zijn **combinaties en varianten** mogelijk, en uiteraard kunnen de functionaliteiten toegepast worden op **om het even welke soort informatie**, niet alleen naam- en adresinformatie.

Voordelen

- **Versnellen van de analysefases**

  In een korte tijdspanne kan in om het even welke context de aanwezigheid aangetoond worden van datakwaliteitsproblemen.

Zoals bijvoorbeeld: het gebrek aan standaardisatie, de aanwezigheid van dubbels, het overtreden van business rules ...

Dit laat toe beter in te schatten wat er dient te gebeuren en hoeveel effort dit zal kosten.

Na de analyse volgt een overlegmoment met de business om een oplossingsstrategie te bepalen. Hierin wordt besproken de problemen al dan niet gedeeltelijk geautomatiseerd met data quality tools aan te pakken.

- **Sneller en beter itereren met business-kenners**
- **Betere ontwikkelingen, lagere maintenance kost**

  Men levert betere ontwikkelingen af, die met datakwaliteitsproblemen rekening houden, en die problemen tijdens de ontwikkeling of in productie vermijden (problemen die anders onvoorzien zouden zijn).

  Dit gebeurt door strategie, methodes en resultaten te laten valideren en door in te spelen op change requests.
- **Accurater inschatten van risico's en required effort**
- **Betere voorbereiding van data-migraties, beter omgaan met de moeilijkheden van data-integratie**

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

Reële data, fuzzy matching en incoherentiedetectie overheen 2 bronnen (L en R) – getoond wordt een cluster van dubbels en de link die zo, zonder key, tussen 2 registers kan worden gelegd.

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

Reële data, adres-cleansing (standaardisatie en matching) – postcode wordt gecorrigeerd, straatnaam wordt gestandaardiseerd, adreselementen worden correct ingedeeld (parsing), gemeentenaam wordt gecorrigeerd, dubbels worden gedetecteerd en georganiseerd in clusters.