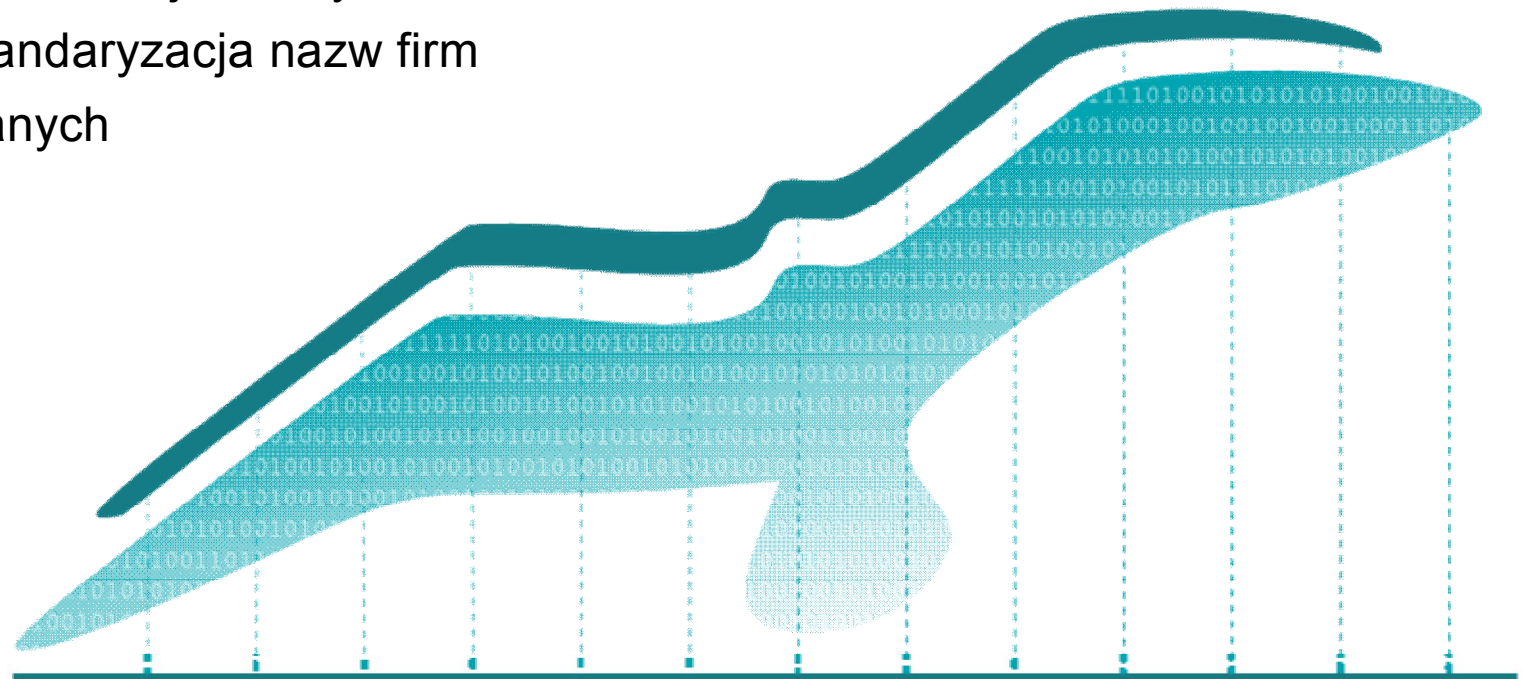


Deduplikacja danych

Zarządzanie jakością danych podstawowych

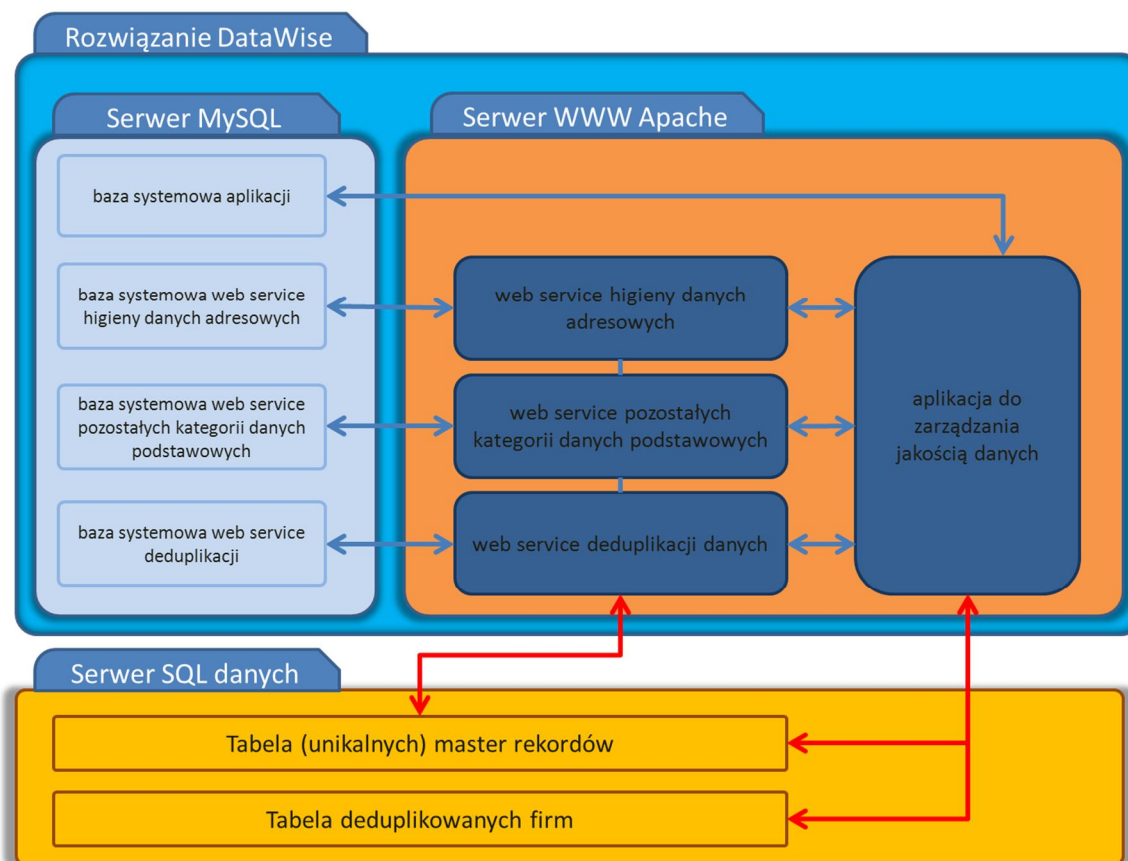
- normalizacja i standaryzacja adresów
- standaryzacja i walidacja identyfikatorów
- podstawowa standaryzacja nazw firm
- deduplikacja danych





Deduplication Center

DataWise udostępnia własne rozwiązanie do deduplikacji danych. Rozwiązanie to ma postać powiązanych ze sobą usług sieciowych (tzw. web services). Poszczególne usługi mogą być wykorzystywane samodzielnie z poziomu własnych aplikacji, a także poprzez interfejs aplikacji webowej.



Web service to standardowa technologia komunikowania się komputerów przez sieć z wykorzystaniem protokołu HTTP do transportu danych oraz składni XML do opisu danych:

- jest niezależna od systemu operacyjnego, dzięki temu ten sam web service jest „widziany” i tak samo „rozumiany” przez różne maszyny (Windows, Linuks czy UNIX)
- jest niezależna od języka programowania, dzięki temu może być wykorzystywany w różnych aplikacjach, np. napisanych w PHP, VisualBasic, C++

Podstawowe korzyści dla firmy:

- jedno rozwiązanie, które może mieć wiele jednoczesnych zastosowań w różnych miejscach (system CRM, serwis WWW, sklep internetowy, call center, hurtownia danych)
- każdy z web service może być wykorzystywany indywidualnie (np. web service higieny danych adresowych przy zgłoszeniach od klientów na stronie www, a web service deduplikacji w systemie CRM)
- wykorzystanie tego samego web service w różnych obszarach de facto wprowadza ten sam standard w tych obszarach, co zapewnia spójność danych
- web service są idealnym narzędziem do automatyzacji zadań, co wydatnie oszczędza zasoby firmy



Web services zarządzania jakością danych

Nasze rozwiązanie pozwala na wykonywanie deduplikacji w następujących wariantach:

- web service higieny danych adresowych – normalizuje, standaryzuje, uzupełnia i waliduje adresy
- web service identyfikatorów NIP, REGON, PESEL – standaryzuje i waliduje oficjalne identyfikatory
- web service nazw firm – dokonuje podstawowej standaryzacji nazw firm
- web service deduplikacji – sprawdza w bazie danych, czy dany rekord już istnieje

Sposoby wykonywania deduplikacji

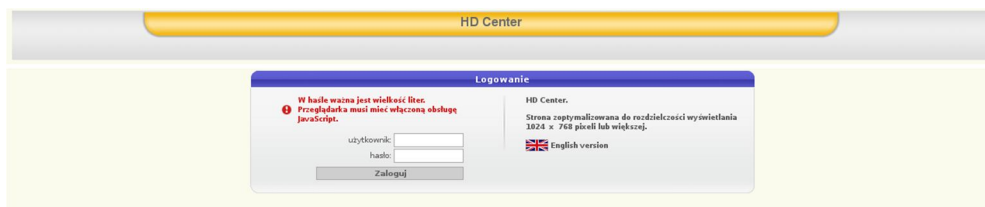
Nasze rozwiązanie pozwala na wykonywanie deduplikacji w następujących wariantach:

- tryb czasu rzeczywistego, co pozwala na bieżąco kontrolować, czy pojedynczy rekord już istnieje w bazie danych
- tryb wsadowy, pozwalający na wykonanie deduplikacji już istniejącej bazy danych. W trybie wsadowym rozróżniamy deduplikację pierwotną (deduplikowana jest cała baza) oraz deduplikację przyrostową (deduplikowane są tylko nowe rekordy danych)

element	deduplikacja w trybie czasu rzeczywistego	deduplikacja w trybie wsadowym
zastosowanie	deduplikacja pojedynczego rekordu danych z istniejącą bazą (sprawdzanie, czy rekord już istnieje w bazie na etapie wprowadzania danych do bazy / rejestracji) w czasie ułamka sekundy	deduplikacja całej istniejącej bazy danych (deduplikacja pierwotna) deduplikacja nowych rekordów względem już istniejącej bazy (deduplikacja przyrostowa)
zwrocane wyniki	jeżeli dla zadanego rekordu znaleziono duplikaty, zwracana jest lista identyfikatorów najbardziej prawdopodobnych duplikatów (co pozwala dalej wyświetlić rzeczywiste dane poszczególnych rekordów i dać możliwość podjęcia decyzji przez operatora czy jest to rzeczywiście duplikat)	do bazy danych zapisywany jest tzw. identyfikator klastra duplikatów (rekordy mające tę samą wartość na klasterze duplikatów uznawane są za duplikaty)

Interfejs aplikacji

Nasze rozwiązanie udostępnia także interfejs aplikacji webowej, pozwalającej użytkownikowi na wykonywanie deduplikacji w trybie wsadowym.



Po zalogowaniu się, użytkownik zobaczy główne menu z poszczególnymi funkcjonalnościami. Korzystanie z aplikacji jest proste polega na korzystaniu z systemu hiperłączy (linków) oraz formularzy do edycji danych i do selekcji danych.

Moduły do wsadowego (batchowego) procesowania danych są uruchamiane przez kliknięcie odpowiedniego hiperłączy. Po uruchomieniu wsadowego procesowania danych przeglądarka www może zostać zamknięta a procesowanie będzie kontynuowane na serwerze aplikacji (procesowanie będzie kontynuowane nawet po zamknięciu komputera użytkownika). W każdej chwili użytkownik może przerwać procesowanie danych.

Aplikacja zawiera także funkcje pozwalające zarządzać wynikami deduplikacji:

- wyświetlanie wszystkie rekordów z danego klastera duplikatów
- zmienianie przypisania indywidualnych rekordów do innego klastera duplikatów
- tworzenie nowego klastera duplikatów
- wyświetlanie logu historii zmian oraz wycofywanie zapisanych zmian
- tworzenie własnych definicji grup klasterów do dodatkowego sprawdzania
- tworzenie indywidualnych raportów i zestawień
- wyświetlanie danych z innych źródeł

Deduplication Center

Wyszukiwanie : Schowek : Tools : Raporty : Log zmian : Ustawienia : Admin : Proces

sprawdzenia -- nie dotyczy id nazwa NIP miasto ulica waniłowa ASM/PH % szukajw: master member pomijaj zrobione Szukaj

LISTA KLASTERÓW MASTER

Menu	ID	Freq	Nazwa	Kod	Miejscowosc	Ulica	NrUlicy	NIP	WoL_3mTotal	DataMin	DataMax	Uwagi	Status	SM
	19594	5	SPOZYWCZO PRZEMYSLOWY RAR ...	81-591	G DYNIA	WANILOWA	6	5860068879	30414	1901-01-01	1901-01-01	0	orygi ...	ANC-03
	97595	3	BODLAK BOLESLAW	81-591	G DYNIA	WANILOWA 6		5831797191	10168	1901-01-01	1901-01-01	0	orygi ...	
	97599	1	PHU 'PRIMA' MALGORZATA BU ...	81-591	G DYNIA	WANILOWA 6		5861002434	800	1901-01-01	1901-01-01	0	orygi ...	

REKORDY NALEŻĄCE DO KLASTERA MASTER @ ID 019594

Menu	Master_ID	Zrodlo	Nazwa	Kod	Miejscowosc	Ulica	NrUlicy	NIP	Uwagi	Status	SM
	19594	TMERT_000050143	SPOZYWCZO PRZEMYSLOWY RARYTAS S.C.	81-591	G DYNIA	WANILOWA	6	5860068879	id klaster ...		ANC-03
	19594	DYST33125	KURA & LEWANDOWSKI	81-591	G DYNIA	WANILOWA 6		5860068879	id klaster ...		
	19594	DYST33126	RARYTAS SKLEP	81-591	G DYNIA	WANILOWA 6		5860068879	id klaster ...		
	19594	DYST187326	SKLEP SPOZYWCZO-PRZEMYSL S.C.'RARYTAS'B.KURA M.LEW ...	81-591	G DYNIA	WANILOWA 6		5860068879	id klaster ...		
	19594	DYST103997	1012880 SKLEP SPOZYWCZO-PRZEMYSLOWY SC RARYTAS BKU ...	81591	G DYNIA	WANILOWA 6		5860068879	id klaster ...		



Co jest potrzebne, aby móc korzystać z serwera do zarządzania jakością danych

Usługi składowe procesu deduplikacji działają jako web service, jest to powszechna technologia wymiany danych pomiędzy komputerami, nawet z różnych platform i systemów operacyjnych. Usługi web service są dostępne w dwóch najpopularniejszych implementacjach:

- XMLRPC
- SOAP

Korzystanie z usług jest możliwe wszędzie tam, gdzie technologia (oprogramowanie) umożliwia łączenie się z web service.

W praktyce możliwe jest nawet udostępnienie usługi dla baz danych napisanych w MS Access czy arkuszach kalkulacyjnych MS Excel, i oczywiście w językach programowania (PHP, .NET, VisualStudio, C, itp.);

Interfejs aplikacji webowej pozwala ponadto na wykonywanie podstawowych operacji importu i eksportu danych w prekonfigurowanych procesach. Do importu i eksportu danych można także wykorzystywać standardowe rozwiązania ETL z wykorzystywanego systemu bazodanowego.



Kiedy i dlaczego stosować

Zastosowania

Nasze web service mają zastosowania w następujących obszarach:

- wprowadzanie nowych rekordów do bazy (różnego rodzaju procesy i formularze rejestracyjne)
- przy projektach, w których istnieje konieczność deduplikacji baz danych
 - deduplikacja baz z różnych źródeł (np. różnych oddziałów firm)
 - poprawa jakości istniejącej bazy, która wcześniej nie była odpowiednio zarządzana od strony deduplikacji
 - pozyskiwanie nowych rekordów z zewnętrznych baz danych
 - łączenie baz z różnych systemów (np. księgowości i CRM) w ramach jednej firmy

Korzyści

- istotne podniesienie jakości bazy danych, co przekłada się wprost na szerszy zakres możliwości wykorzystywania bazy
- umożliwia łączenie informacji z różnych systemów bazodanowych, co daje bardziej kompletny obraz o podmiocie (tzw. ogląd 360°)
- pozwala na rzeczywistą integrację danych do jednego spójnego systemu bazodanowego

Modele wykorzystywania usługi

Ze względu na fakt, że deduplikacja danych z definicji odnosi się do własnych zasobów bazodanowych w firmie, możliwe są w praktyce dwa modele wykorzystywania usługi:

- wdrożenie kompletnego rozwiązania na serwerze firmy.
DataWise dostarcza firmie gotowy serwer, szkoli pracowników, zapewnia usługę pomocy i opieki technicznej.
- wykonanie usługi w trybie projektowym.
DataWise wykonuje usługę deduplikacji na przekazanych przez firmę danych (zwykle w postaci plików tekstowych). Po wykonaniu deduplikacji dane są zwracane w ustalonym formacie (zwykle w postaci plików tekstowych).



Dodatkowe informacje w Internecie

Dodatkowe informacje o rozwiązaniach DataWise do zarządzania jakością danych można znaleźć na stronie <http://datawise.pl>

Na stronie można skorzystać z formularza pozwalającego na testowanie usługi online <http://datawise.pl/demo/deduplikacja-danych>

Kontakt

Krzysztof Pędzich

GSM 0-501-725-574

email k.pedzich@datawise.pl

Marek Turlejski

GSM 0-501-099-698

email m.turlejski@datawise.pl