



ZARZĄDZANIE JAKOŚCIĄ DANYCH (DATA QUALITY)

Profilowanie danych (Data Profiling)

StatConsulting oferuje usługę profilowania danych, która polega na ich eksploracji pod kątem wykrywania zjawisk obniżających jakość danych oraz mogących mieć wpływ na wyniki przeprowadzanych analiz.

Profilowanie jest etapem prac analitycznych poprzedzającym większość zadań związanych z Data Quality. Jednym z analizowanych w trakcie profilowania zagadnień jest np. weryfikacja, czy dane wymagają uspoźnienia poprzez proces standaryzacji.

W oferowanych usługach wykorzystujemy własne narzędzia do profilowania i eksploracji danych, które umożliwiają m.in. sprawdzanie poprawności technicznej (podstawowe statystyki danych, braki danych, testy formatu danych itp.) oraz poprawności merytorycznej (testy spójności danych, zgodność ze słownikami, wartości odstające, nietypowe itp.). W profilowaniu wykorzystujemy również metody Data Mining, które umożliwiają odkrywanie nieoczywistych zjawisk i charakterystyk w danych.

W wyniku profilowania klient otrzymuje raport z informacją o poziomie jakości danych oraz ze wskazaniem usług niezbędnych do podniesienia tej jakości.

Czyszczenie danych (Data Cleansing)

StatConsulting oferuje także usługi czyszczenia danych. Czyszczenie danych rozwiązuje znane lub poznane w trakcie profilowania problemy jakości danych i jest najpowszechniej kojarzonym etapem Data Quality. Nasze rozwiązania w ramach czyszczenia danych obejmują zarówno proste w realizacji zadania - jakim jest np. usunięcie niepotrzebnych białych znaków - jak i te bardziej zaawansowane: parsing, standaryzacja i deduplikacja.

W rezultacie procesu czyszczenia zwiększa się poprawność i spójność posiadanych przez firmę danych.

Parsowanie danych (Parsing)

Usługa parsowania danych polega na identyfikacji i wydobyciu określonych elementów danych z pól o nieformalnej strukturze, w oparciu o znaczenie danych i kontekst, np. wydobycie informacji o imionach klienta z pola 'Imię i Nazwisko', lub informacji o ulicy, numerze domu, numerze mieszkania, kodzie pocztowym z pola 'Adres'.

Rozwiązania StatConsulting posiadają specjalistyczne algorytmy do parsowania, a w sytuacjach nietypowych istnieje możliwość dopasowania ich do konkretnego zadania parsowania.

W rezultacie procesu parsowania, klient otrzymuje bazę, w której dotychczasowe wieloelementowe pola zostają podzielone na oddzielne kolumny gotowe do użycia w procesach biznesowych i analitycznych.

Standaryzacja danych (Standardization)

Oferowana przez StatConsulting usługa standaryzacji ma na celu zamianę wielu różnych form wyrażenia na jedną formę.

Standaryzacja jest niezbędna między innymi dla poprawności różnego rodzaju raportów np. wartości zakupów według nazw kontrahentów. Standaryzację stosuje się w sytuacji, gdy w danych ta sama informacja zapisana jest na wiele różnych sposobów, np. w bazie kontrahentów występują 3 różne sposoby opisu dla tej samej firmy: StatConsulting, StatConsulting sp z o.o, StatConsulting Sp. z o.o. Standaryzacja obejmuje zarówno proste ujednocnianie formatów (kody pocztowe, numery telefonu) jak również bardziej złożone uspoźnianie danych przy użyciu tabel synonimów, słowników czy predefiniowanych reguł (ulice, adresy email, numery PESEL, kody PKD).

StatConsulting posiada między innymi słowniki wykorzystywane do standaryzacji wartości w danych demograficznych (nazwy miast, imiona, adres), tworzy również słowniki na podstawie danych klienta.

Wynikiem oferowanej usługi standaryzacji będzie baza danych, w której będą zdefiniowane i zachowane standardy zapisu różnego typu informacji. Dzięki temu zwiększy się wiarygodność statystyk bazujących na takich danych, przez co zwiększy się zaufanie do raportów, a podejmowane na ich podstawie decyzje będą bardziej precyzyjne.



Deduplikacja danych (Deduplication)

Deduplikacja jest stosowana w celu eliminacji wielokrotnych wpisów w bazie danych. Jako duplikaty rozumiemy rekordy odnoszące się do tego samego obiektu (np. klienta indywidualnego, firmy, transakcji zakupowych). Duplikaty mogą powstawać w wyniku różnej interpretacji danych lub pomyłek przy wprowadzaniu danych i mogą stanowić od kilku do kilkunastu procent zawartości, co wpływa na podwyższenie kosztów np. każdej wysyłki do klienta.

Oferujemy rozwiązania eliminacji duplikatów, które w efekcie pozwalają m.in. zmniejszyć koszty kampanii marketingowych i działań CRM.

W ramach deduplikacji oferujemy:

- identyfikację duplikatów – wskazanie i wyodrębnienie zbioru zduplikowanych wpisów, a także wskazanie zbioru potencjalnych duplikatów,
- konsolidację bazy – przy współpracy z klientem opracowanie i wdrożenie reguł konsolidacji.

W procesie identyfikacji duplikatów wykorzystujemy dwa rodzaje matchingu:

- matching równościowy – porównywanie rekordów metodą idealnej zgodności,
- matching probabilistyczny – oparty na regułach porównywania z dopuszczeniem niewielkich różnic.

Wykorzystanie matchingu probabilistycznego umożliwia:

- dostosowanie się do specyfiki konkretnych danych (np. uwzględnienie polskich standardów zapisu numerów telefonów, adresów),
- zwiększenie ilości wykrywanych duplikatów,
- zachowanie wysokiej precyzji.

W wyniku deduplikacji klient otrzymuje oczyszczoną z duplikatów bazę, dzięki której podejmowane działania m.in. marketingowe będą efektywniejsze a analizy bardziej wiarygodne.

Wzbogacanie danych (Data Enrichment)

Wzbogacanie przez łączenie zewnętrznych źródeł

Nasza oferta obejmuje usługę wzbogacania danych z wykorzystaniem m.in. własnych lub zewnętrznych baz danych. Łączenie zewnętrznych źródeł stosuje się do takich celów jak np.:

- dołączanie dodatkowych informacji o klientach, np. wykorzystanie zewnętrznej bazy danych o konsumentach do poznania pewnych cech klientów,
- poznawanie części wspólnej różnych baz klientów.

Rekordy mogą występować w tej samej lub w wielu różnych bazach danych i nie muszą mieć idealnie spójnego formatu.

Dzięki łączeniu danych z różnych źródeł można uzyskać pełniejszą charakterystykę klientów (dodatkowe informacje o zachowaniach klientów), zwiększyć wartość raportów (agregacja danych według nowej, dołączonej zmiennej) i w efekcie podejmować bardziej uzasadnione decyzje.

Wykrywanie gospodarstw domowych (Householding)

W ramach projektów Data Quality StatConsulting oferuje również usługi z zakresu Householding rozumianego jako wykrywanie połączeń między klientami, takich jak przynależność do tego samego gospodarstwa domowego lub firmy. Usługa ta pozwala na odkrycie i wykorzystanie niedostępnych do tej pory informacji o strukturze posiadanych danych. Wynikiem Householdingu jest baza zawierająca klientów posegregowanych tak, że działania marketingowe mogą być adresowane do gospodarstwa domowego zamiast do poszczególnych jego członków. W efekcie zmniejsza to koszty np. wysyłki pocztowej.

Przewidywanie wartości kluczowych zmiennych

Oferujemy użycie technik Data Mining i metod statystycznych w przewidywaniu (wypełnianiu kontekstowym) kluczowych zmiennych. Można je stosować do realizacji specyficznych zadań, m.in.:

- wypełniania wartości brakujących,
- wykrywania i korekty nieprawidłowych wartości,





- predykcji interesującej wartości.

Usługa ta jest specyficzna dla rozwiązywanego zagadnienia, wspiera m.in. podejmowanie decyzji poprzez dostarczenie wcześniej ukrytych informacji. Przykładowym zastosowaniem jest wykorzystanie metod Data Mining do predykcji klasy zamożności klienta na podstawie dokonywanych przez niego transakcji zakupowych. Innym przykładem jest wskazanie nieprawidłowych danych, podanych błędnie, celowo przez klienta, np. zawyżenie lub zaniżenie dochodów lub wieku.

Wykorzystanie słowników

Niezależnie od wybranej usługi w zadaniach dotyczących Data Quality może być niezbędne wykorzystanie słowników. Oferujemy wykorzystanie słowników do :

- konwersji danych zmieniających się w czasie ze starego formatu do nowego, np. tłumaczenie kodu klasyfikacji działalności EKD i PKD na PKD2004,
- dodawania nowych informacji do istniejącej bazy, np. forma wołacza dla imienia przy adresowaniu przesyłek,
- weryfikacji poprawności wartości, np. weryfikacja przynależności miasta do województwa czy weryfikacja zakresu przyjmowanych wartości przez zmienne kodowane,
- standaryzacji pól takich jak imiona, nazwy ulic i miejscowości.

Przykładem może być dołączenie liczby mieszkańców danego regionu na podstawie kodu pocztowego. W rezultacie otrzymujemy nową informację - tutaj liczbę mieszkańców, którą można wykorzystywać w celach operacyjnych i analitycznych.

W naszych projektach dysponujemy własnymi słownikami (np. słownikiem miejscowości, kodów pocztowych i ulic), umożliwiamy utworzenie słownika na podstawie danych klienta oraz wykorzystanie lub weryfikację słownika dostarczonego przez klienta.

Kontrola i monitoring jakości danych (Data Quality Control)

W trakcie realizowania usług poprawiających jakość danych, często celem jest osiągnięcie i utrzymywanie pewnego poziomu jakości danych w dłuższym okresie czasu.

Ponieważ dane podlegają ciągłym zmianom, oferujemy możliwość zautomatyzowania wybranych procesów Data Quality i ich cykliczne uruchamianie. Przykładowo, wszystkie nowe wpisy w bazie danych podlegają automatycznej standaryzacji.

W celu nadzoru automatycznego procesu, jego przebieg i wyniki działania są raportowane i monitorowane.

Oprócz procesów czyszczenia, nasze rozwiązania obejmują również monitoring wskaźników jakości danych. Dzięki temu klient otrzymuje aktualną informację o poziomie jakości danych oraz możliwość odniesienia bieżącej sytuacji do przeszłości.

Oprogramowanie

Firma StatConsulting oferuje także [aplikację do zarządzania jakością danych \(Data Quality\) - StatDQ](#).

