TRANSPARENT DATA
DATA SOFTWARE HOUSE

Case Study Data Services: Parsowanie danych z tabel z plików PDF

Wstęp

Kilka słów o projekcie

Nasz Klient, podobnie jak wiele firm pobierających informacje biznesowe z rozmaitych, często “trudnych” źródeł, stanął przed problemem przekształcenia danych tabelarycznych z PDF w sposób, który rzeczywiście będzie efektywny i rzetelny. Jego poprzedni parser danych tabelarycznych nie sprawdzał się dobrze w praktyce, zwracając ilość błędów na poziomie 19,9%.

Format PDF nie jest przyjazny przetwarzaniu maszynowemu - stanowi połączenie pliku binarnego i tekstowego, któremu daleko do miana ustrukturyzowanych danych. W przypadku zwykłego tekstu parsowanie danych z plików PDF bywa często po prostu czasochłonne. Gdy jednak dokumenty udostępniane w tym formacie mają strukturę tabel, niejednokrotnie rozciągniętych na kilka stron, z polami wypełnionymi nieprzewidywalnej długości tekstem, czasem literówkami, parsowanie danych staje się prawdziwym wyzwaniem. Ilość błędów parsera przekłada się proporcjonalnie na koszty pozyskania, przetwarzania, a także dalszego wykorzystania pozyskanych informacji.

Jako ekspert w dziedzinie obróbki danych podjęliśmy się dla Klienta budowy całkowicie nowego parsera danych tabelarycznych z plików PDF.

Rodzaj wykonanych usług data services: parsowanie danych tabelarycznych z plików PDF
Rok: 2021
Czas trwania projektu: 1 miesiąc

Potrzeby biznesowe

Klient używał parsera plików PDF wyciągającego dane z tabel, a wyciągnięte w ten sposób dane były później używane w wewnętrznej aplikacji webowej, wspierającej zarządzanie ryzykiem i szybką kwalifikację leadów B2B zgodnie z przyjętą procedurą KYC/KYB. Wysoki poziom błędów występujących w procesie parsowania przekładał się na znaczne koszty procesu biznesowego. Niekompletne lub mylące informacje generowały koszty nie tylko na poziomie samego ręcznego poprawiania danych i utrzymywania higieny bazy, ale ceną były również błędne decyzje podejmowane w oparciu o nie.

Potrzebą Klienta było więc znaczne poprawienie jakości procesu parsowania. Na współpracę z naszym data software house zdecydował się z uwagi na duże doświadczenie Transparent Data w technologicznej obróbce i agregowaniu ogromnych zbiorów danych gospodarczych.

Czego oczekiwał Klient od nowego parsera danych tabelarycznych?

Opracowanie nowego sposobu parsowania danych z tabel z PDF miało z założenia zminimalizować błędy do maksymalnie 4%. Klient liczył na to, że całym zadaniem zajmiemy się kompleksowo - od przeprowadzenia analizy danych i wyśledzenia błędów oraz ich kwalifikacji po napisanie nowego parsera, którego zadaniem jest konwersja danych w formacie PDF na dane o formacie JSON według określonego schematu. Podjęliśmy się również pomocy Klientowi w implementacji nowego rozwiązania w jego wewnętrznej aplikacji.

Rozwiązanie

Efekty nowego sposobu parsowania danych z tabel z plików PDF

- 63 x mniej wszystkich błędów podczas parsowania danych (ilość błędów wynosi średnio zaledwie 0,31%)
- Całkowita eliminacja takich powtarzalnych błędów, jak literówki czy źle opisane numery PESEL
- Średni czas parsowania pojedynczego dokumentu PDF jest 7 x krótszy

Jaką pracę wykonała Transparent Data?

Pierwszym etapem był audyt metodyki starego parsera Klienta i analiza danych występujących w plikach PDF w celu wychwycenia wszystkich możliwych rodzajów błędów. W ten sposób zlokalizowaliśmy, że 78% wszystkich błędów to cyfry występujące w imionach oraz źle opisane numery identyfikacyjne PESEL. Druga duża pula błędów wynikała z długości przechowywanych w tabelach danych i ucinaniu danych zawartych w kolejnych liniach.

Jednym z powodów występowania takich problemów było nietrafne podejście, opierające się na zastosowaniu popularnego narzędzia pdftotext do konwertowania plików PDF na pliki tekstowe. Pierwotny parser danych tabelarycznych z PDF opierał się na sczytywaniu informacji między liniami, czyli na licznych dopasowaniach tekstu poprzez wyrażenia regularne, swego rodzaju

“kotwice”. Występowanie tych “kotwic” było oczekiwane i poszukiwane. Znaleziony fragment tekstu pomiędzy dwiema “kotwicami” był wówczas pobierany i zapisywany. Takie podejście rodziło wiele problemów, w szczególności w przypadku nieuniknionych w analizowanym tekście literówek lub nietypowego układu wygenerowanej tabeli. W takich sytuacjach wyrażenia regularne nagminnie zwracały puste dopasowanie albo dopasowanie zawierające nadmiarowe i niepożądane znaki. Ostatecznie, często tak zniekształcone albo niepełne dane trafiały wprost do aplikacji biznesowej Klienta, obniżając jej wiarygodność i wartość użytkową.

Nowy parser danych z tabel z plików PDF stworzyliśmy w technologii, która umożliwia bardziej **analityczne podejście do rozwiązywanego problemu i równoległe przetwarzanie** (tj. w tym samym czasie) każdej strony dokumentu. Zamiast opierać się na wyrażeniach regularnych z wieloma instrukcjami warunkowymi, nowe rozwiązanie pozwala na dostarczenie schematu konfiguracyjnego (tzw. JSON Schema) do parsowanego pliku. Ustala pozycję każdego prostokąta i każdego wpisanego w niego znaku. Znaki łączy w zdania i zapisuje w komórce, komórki łączy w wiersze, wiersze w tabele i jeśli jest taka potrzeba, to scala tabele, które bywają rozdzielone na kilka stron.

Dzięki nowemu podejściu parser PDF analizuje tabele względem ustalonych schematów, wykrywając liczbę wierszy, kolumn i zagnieżdżonych podtabel. W efekcie tego działania produkuje ustrukturyzowaną formę treści tabel w całym analizowanym pliku. Co istotne, schemat JSON jest łatwy do utworzenia nawet dla osób nietechnicznych i przeprowadziliśmy z Klientem warsztaty, podczas których wspólnie utworzyliśmy schematy dostosowane do potrzeb jego procesów.

Poza dodatkowym zyskiem wydajnościowym, nowe rozwiązanie parsowania danych jest bardziej uniwersalne, generyczne i konfigurowalne, a także łatwiejsze do zaadaptowania w innych miejscach, które operują na wartościach tabelarycznych w PDF. Akceptujące określony margines błędu (tzw. odległość Levenshteina) algorytmy przetwarzania tekstu umożliwiają poprawne dopasowanie tabel na bazie ich nagłówków, nawet gdy zawierają one drobne błędy czy literówki.

Największe wyzwania

- Największym wyzwaniem była szeroka analiza danych zawartych w plikach PDF, tak aby stworzyć własną typologię 60 różnych, powtarzalnych schematów danych. W oryginalnych dokumentach widoczny był bowiem:
    - Brak wewnętrznej struktury, na której można oprzeć mechanizmy parsujące dane (PDF jest z założenia formatem służącym do prezentacji treści tekstowo-graficznych - nie jest dobrym sposobem na wymianę danych pomiędzy systemami teleinformatycznymi)
    - Brak reguł dla powtarzalnego złamania stron i tekstu w tabelach
    - Brak reguł z wypełnieniem poszczególnych komórek (np. brak wartości w wierszu lub inny niż spodziewany typ danych)
- Dodatkową komplikacją był fakt, że tabele w pliku PDF były generowane automatycznie z nieznanego oprogramowania, co w pewnych sytuacjach rodziło nietypowe i nieprzewidziane drobne błędy wizualne. Np. gdy tabela nie mieściła się na jednej stronie, zostawała rozciągnięta na wiele kolejnych stron dokumentu. Te wizualne niedoskonałości wprost przekładały się na proces dopasowania tekstu, obniżając jakość dostarczanych danych. Jednocześnie będąc dynamicznie generowaną tabelą, nie posiadała jednorodnej struktury. Zarówno liczba wierszy tabeli, jak i liczba zagnieżdżonych w niej podtabel była zmienna i dynamiczna w zależności od dostarczonego pliku
- Poza samą zmianą sposobu parsowania danych projekt pociągnął za sobą stworzenie całkowicie nowej bazy danych, w tym opracowania nowej struktury danych oraz integracji starych systemów Klienta z nową bazą. Wymagało to kompleksowych działań w wielu obszarach.

Wykorzystane technologie

- Go (Golang)
- PHP 8.0/8.1
- RabbitMQ

- Redis
- Laravel (Lumen)
- Kubernetes
- Docker
- Google Cloud Platform