**Opis projektu**

Motywacja:

W ramach prowadzonych analiz, w tym zadań dot. przygotowania gotowych raportów dla kadry kierowniczej otrzymuję dane z różnych źródeł. Dane te są generowane z lokalnych aplikacji, baz danych, systemów ERP, CRM, danych z sieci Internet, plików płaskich. Dane te są najczęściej w różnych strukturach i formatach. Takie „surowe” dane są najczęściej „zabrudzone” i „zaszumione” i nie nadają się do bezpośredniego zastosowania do analiz czy raportowania.

W celu przygotowania danych „gotowych” do użycia w większości organizacji dane przechodzą przez proces ETL (wyodrębnianie, przekształcanie i ładowanie), zanim zostaną udostępnione do raportowania. Podczas procesu ETL dane są pobierane ze źródła danych, a następnie przekształcane, weryfikowane, standaryzowane, korygowane, sprawdzane pod kątem jakości i ostatecznie ładowane do repozytorium danych - takiego jak zbiorcza baza danych lub hurtownia danych - gdzie są usprawnione do analizy i raportowania.

Niżej wskazany przypadek opisuje możliwość zastosowania szybkiego, skutecznego i darmowego narzędzia do procesów ETL - Power Query. To narzędzie sprawdzi się w większości przypadków również dla dużych woluminów danych.

Opis przypadku (Case study):

Firma PPHU „ZYZQ” prowadzi działalność w zakresie produkcji i sprzedaży mebli. W 2020 roku dział KADR w związku z planowaną reorganizacją w firmie tj. podział dotychczasowych komórek organizacyjnych i stworzenie nowych działów powstałych z wyodrębnienia już funkcjonujących otrzymał zadanie od zarządu do przedstawiania raportu w zakresie oceny wykorzystania zasobów kadrowych i ewentualnych zmian w ww. zakresie w związku z dostosowaniem do nowej struktury organizacyjnej firmy.

Cele przedstawione dla działu KADR:

- przygotować zestawienie pracowników firmy z obecnym przydzieleniem ich do poszczególnych działów wraz z danymi kadrowo-płacowymi oraz danymi pochodzącymi z badania ankietowego wśród wszystkich pracowników;

- zebrać informacje od wszystkich pracowników w zakresie

\* przyszłych potrzeb szkoleniowych

\* chęci pracownika do zmiany stanowiska/działu

\* ogólnej oceny zadowolenia pracownika z pracy

\* zadowolenia z zarobków

Przedstawić w formie tabelarycznej umożliwiającej prowadzenie dalszych analiz i tworzenia raportów.

W opisywanym przypadku informacje niezbędne do przygotowania zestawienia danych dla zarządu będą pochodzić z wielu różnych źródeł danych (zarówno ustrukturyzowanych, jak   
i nieustrukturyzowanych). Zatem niezbędne jest z rozwiązania, które będzie w stanie gromadzić   
i wizualizować dane oraz mieć możliwość udostępniania wyników zespołowi w jasny i zwięzły sposób. Oczywiście, że część zadań ETL można wykonać bezpośrednio w programie Excel, jednakże proces ETL w programie Excel jest zwykle czasochłonnym, ręcznym procesem, który nie jest łatwy   
do zautomatyzowania, w znacznej części musiałby korzystać z kodu VBA. Z uwagi na powyższe   
do rozwiązania ww. problemu zastosowano dodatek do MS EXCEL - Power Query.

Opis procesu:

***Etap - wyodrębnianie (Extract)* - przygotowanie danych**

a) pozyskanie danych z różnych źródeł

Dane , które posłużą do skonstruowania tabeli danych pochodzić będą z różnych źródeł:

* Dane z działu KADR – generowane w pliku \*.csv pochodzą z lokalnego programu do zarzadzania kadrami
* Dane z działy PŁAC – generowane w pliku \*.xlsx pochodzą z programu finansowo-płacowego
* Dane z Ankiet przeprowadzone wśród wszystkich pracowników (zbierane w formie papierowej ankiety w ciągu 2 tygodni – zwrot ankiet - 100%) – następnie dane te były ręcznie przenoszone do arkusza kalkulacyjnego \*.xlsx.

b) łączenie danych

- dane załadowane będą do Power Query jako 3 niezależne tabele zawierające pobrane ww. dane,  
 które następnie zostaną poddane odpowiednim przekształceniom.

***Etap - przekształcenie (Transform)***

* przekształcanie danych zawartych w ww. tabelach poprzez:

Uporządkowanie danych

* 1. przekształcenie danych do postaci tabelarycznej
  2. poprawienie nieprawidłowych wpisów

Czyszczenie danych

* 1. Usuwanie lub poprawienie błędnych danych
  2. Usuwanie duplikatów

***Etap - ładowanie (Load)***

Oczyszczone i uporządkowane dane zostaną załadowane do jednej tabeli, kluczem do połączenia danych w ww. przypadku będzie unikatowy *„kod pracownika”*

Schemat procesu:

*Źródła danych Power Query ostateczna tabela danych*

*Raport.xlsx*

Dane z działu KADR

*Kadry.xlsx*

Dane z działu PŁAC

*Płace.csv*

***Extract Transform Load***

Dane z Ankiet

*Ankiety.xlsx*

Dane dot. Działów

*Działy.pdf*

Po dokonaniu pobrania danych z wyżej opisanych źródeł zidentyfikowano, wiele niespójności, niezgodności i błędów w zawartych danych.

w pliku Kadry:

* w kolumnie „dział” zamiast nazwy działu jest jego numeryczne oznaczenie od 1 do 13;
* wartości kolumny dział podzielone między kilka tabel (kadry - nr działu, dział - nazwa działu)
* brak danych dotyczących stażu pracy w firmie (jest data zatrudnienia;
* brak danych dotyczących wieku pracownika (jest rok urodzenia);
* wartości roku urodzenia i płci znajdują się w jednej kolumnie.

w pliku Płace:

* jako nagłówki w tabeli znajdują się nieoznaczone nazwy kolumn;
* nagłówki w tabeli rozpoczynają się od 2 wiersza;
* wartości wynagrodzenia zasadniczego znajdują się w 2 kolumnach (pełne wartości rozdzielone z groszami);
* zakodowane dane (0,1) w kolumnie dodatek zadaniowy;
* zakodowane dane (0,1) w kolumnie dodatek rodzinny.

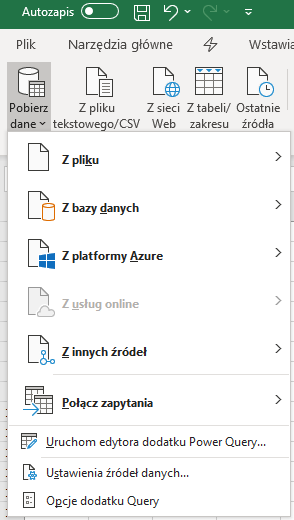
w pliku Ankieta:

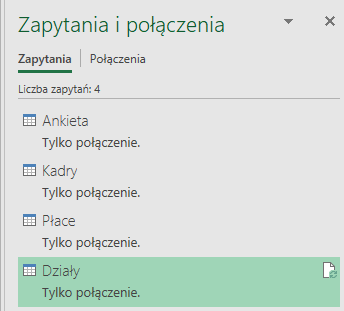
* zidentyfikowano zduplikowane dane;
* różne oznaczenia odpowiedzi (tak/nie) w kolumnie „chęć uczestnictwa w szkoleniach”
* literówki, „czeskie błędy” w kolumnie” preferowana forma szkoleń”
* niespójne dane w przypadku odpowiedzi „nie” (chęć zmiany działu) i jednocześnie uzupełnienie danych (preferowany dział).

w pliku Działy:

- plik zapisany w pliku \*. pdf

1 etap - pobrano i załadowano do Power Query poszczególne pliki źródłowe do pliku Rapoty.xlsx – jako połączenia.





2 Etap – Transformacja poszczególnych plików

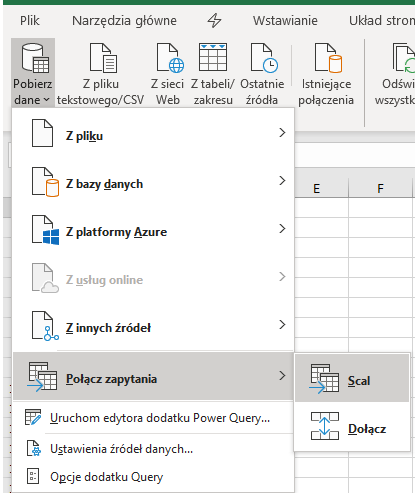
W Power Query dokonano przekształceń i poprawienia danych w taki sposób, by wyeliminować opisane powyższej zidentyfikowane nieprawidłowości.

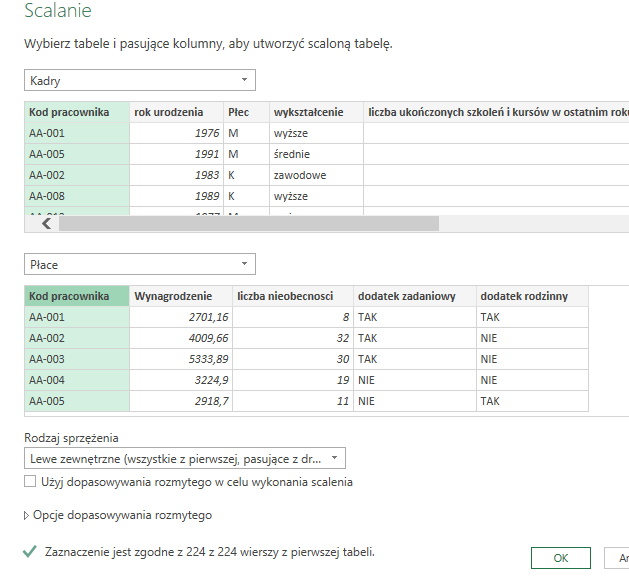
Rozwiązanie ww. nieprawidłowości w Power Query:

|  |  |
| --- | --- |
| **Zidentyfikowane nieprawidłowości** | **Propozycja rozwiązania w Power Query.** |
| **w pliku Kadry:** |  |
| w kolumnie „dział” zamiast nazwy działu jest jego numeryczne oznaczenie od 1 do 13; | Scalono tabele kadry (wg nr działu) z tabelą działy i uzyskano nazwy działu. |
| wartości kolumny dział podzielone między kilka tabel (kadry - nr działu, dział - nazwa działu) | Scalono tabele kadry (wg nr działu) z tabelą działy i uzyskano nazwy działu. |
| brak danych dotyczących stażu pracy w firmie (jest data zatrudnienia; | Dodano nową kolumnę niestandardową (staż) i wstawiono wyliczenie:  „2020-Date.Year([data zatrudnienia])” która wylicza liczbę lat pracy w firmie. |
| brak danych dotyczących wieku pracownika (jest rok urodzenia); | Dodano nową kolumnę niestandardową (wiek) i wstawiono wyliczenie:  „2020-[rok urodzenia]” która wylicza liczbę lat pracownika |
| wartości roku urodzenia i płci znajdują się w jednej kolumnie. | Użyto funkcji *„Podziel kolumny” wg ogranicznika i zmieniono nazwy kolumn na rok urodzenia i płeć* |
| **w pliku Płace:** |  |
| jako nagłówki w tabeli znajdują się nieoznaczone nazwy kolumn; | Usunięto 1 wiersz, |
| nagłówki w tabeli rozpoczynają się od 2 wiersza; | Użyto 1 wiersza jako nagłówków |
| wartości wynagrodzenia zasadniczego znajdują się w 2 kolumnach (pełne wartości rozdzielone z groszami); | Scalono kolumny „wynagrodzenie zadadnicze1 i „wynagrodzenie zadadnicze2”, jako separator wskazano przecinek, zmieniono format na dziesiętny |
| zakodowane dane (0,1) w kolumnie dodatek zadaniowy; | Zmieniono format danych na tekstowy, zamieniono wartości z 0 na NIE oraz z 1 na TAK |
| zakodowane dane (0,1) w kolumnie dodatek rodzinny. | Zmieniono format danych na tekstowy, zamieniono wartości z 0 na NIE oraz z 1 na TAK |
| **w pliku Ankieta:** |  |
| zidentyfikowano zduplikowane dane; | Usunięto duplikaty |
| różne oznaczenia odpowiedzi (tak/nie) w kolumnie „chęć uczestnictwa w szkoleniach” | Zamieniono różne warianty odpowiedzi na TAK i NIE |
| literówki, „czeskie błędy” w kolumnie” preferowana forma szkoleń” | Poprawiono błędy |
| niespójne dane w przypadku odpowiedzi „nie” (chęć zmiany działu) i jednocześnie uzupełnienie danych (preferowany dział). | Usunięto błędny wpis |

3 Etap - ładowanie przekształconych plików do postaci „jednotabelowej”

Oczyszczone i uporządkowane dane zostaną załadowane do jednej tabeli, kluczem do połączenia danych w ww. przypadku będzie unikatowy *„kod pracownika”*

**



W ten sposób uzyskano raport łączący przekształcone pliki w jedną tabelę.

Szczegóły w pliku „Raport.xlsx” stanowiący podstawę do dalszych działań analitycznych.