POLITECHNIKA
OPOLSKA

PRZEGLĄD NAUK STOSOWANYCH

pod redakcją
Mariusza Rząsy

nr 19

Wydział Ekonomii i Zarządzania
Opole, 2018

Przegląd Nauk Stosowanych

Nr 19

ISSN 2353-8899

Przegląd Nauk Stosowanych Nr 19 (2)

Redakcja: Mariusz R. Rząsa

Wszystkie artykuły zostały ocenione przez dwóch niezależnych recenzentów

All contributions have been reviewed by two independent reviewers

Komitet Naukowy czasopisma:

dr hab. Mariusz Zieliński (przewodniczący)
dr inż. Małgorzata Adamska, dr hab. Maria Bernat, dr Ewa Golbik-Madej,
dr Anna Jasińska-Biliczak, dr hab. Izabela Jonek-Kowalska, dr inż. Brygida Klemens,
dr hab. Barbara Kryk, dr Małgorzata Król, dr hab. Aleksandra Kuzior,
prof. dr hab. Krzysztof Malik, dr hab. Mirosława Michalska-Suchanek, Roland Moraru,
PhD. Prof. (Rumunia), doc. PhDr. Michal Oláh PhD (Słowacja),
Volodymyr O. Onyshchenko, Ph.D. Prof. (Ukraina), dr hab. Kazimierz Rędziński,
dr Alina Rydzewska, dr hab. Brygida Solga, dr inż. Marzena Szewczuk-Stępnień,
dr hab. Urszula Szuścik, doc. PhDr. ThDr. Pavol Tománek, PhD (Słowacja), PhDr. Jiří Tuma,
PhD (Republika Czeska), dr hab. inż. Janusz Wielki

Komitet Redakcyjny:

dr hab. Mariusz Zieliński (przewodniczący)

dr inż. Małgorzata Adamska, dr hab. Maria Bernat, prof. dr hab. Krzysztof Malik,

dr hab. inż. Janusz Wielki, dr inż. Magdalena Ciesielska (sekretarz)

Recenzenci:

Przemysław Adamkiewicz, Artur Andruszkiewicz, Robert Banasiak, Agnieszka Dornfeld Kmak,
Tadeusz Dyr, Robert Hanus, Mariusz R. Rząsa, Radosław Wajman, Józef Wiora, Mariusz Zieliński

Copyright by Politechnika Opolska 2018

Projekt okładki: Krzysztof Kasza

Opracowanie graficzne: Oficyna Wydawnicza Politechniki Opolskiej

Wydanie I, 2018 r.

ISSN 2353-8899

Spis treści

Paweł CYBULSKI
SŁOWO WSTĘPNE 5

Justyna BIOŁY-KOBYLAŃSKA
KONFERENCJIA „PRAKTYCZNE ASPEKTY I MOŻLIWOŚCI WYKORZYSTANIA POTENCJAŁU NAUKOWO-BADAWCZEGO ORAZ TRANSFER WIEDZY POMIĘDZY SEKTOREM NAUKI, A JEDNOSTKAMI KRAJOWEJ ADMINISTRACJI SKARBOWEJ" 7

Krzysztof MALIK, Barbara BĘTKOWSKA-CELA, Agnieszka DORNFELD-KMAK
MODEL WSPÓŁPRACY IZBY ADMINISTRACJI SKARBOWEJ W OPOLU Z POLITECHNIKĄ OPOLSKĄ 17

Krzysztof MALIK, Barbara BĘTKOWSKA-CELA, Agnieszka DORNFELD-KMAK
CZY WARTO SKONSOLIDOWAĆ SYSTEMY ZARZĄDZANIA? ANALIZA DOKONANA W OPARCIU O SYSTEMY ZARZĄDZANIA W KRAJOWEJ ADMINISTRACJI SKARBOWEJ 27

Robert EHRMANN
LABORATORIA KRAJOWEJ ADMINISTRACJI SKARBOWEJ 37

Piotr KRACZMAR, Mariusz R. RZĄSA
PROBLEMATYKA POBORU PRÓBEK W CYSTERNACH PRZEWOŻĄCYCH MATERIAŁY PODLEGAJĄCE KONTROLI CELNO-SKARBOWEJ 45

Mariusz R. RZĄSA
WPŁYW LICZBY PRÓBEK NA ODCHYLENIE UŚREDNIONEGO PARAMETRU CIECZY POBRANEJ Z CYSTERNY 55

Przemysław KRAWCZYK, Przemysław MISIURSKI
ANALIZA DANYCH PODATKOWYCH – ZARYS PROBLEMU 61

Wojciech ZIMOCH
NARZĘDZIA INFORMATYKI ŚLEDCZEJ W SŁUŻBIE ZWALCZANIA PRZESTĘPCZOŚCI EKONOMICZNEJ 73

Rafał KOKOT, Tomasz TURBA
ZARYS HISTORYCZNY SIECI DARKNET ORAZ ASPEKTY LEGALNEGO I NIELEGALNEGO WYKORZYSTANIA TECHNOLOGII TOR 83

Mariusz R. RZĄSA, Wojciech GĘSIKOWSKI
TECHNIKI KOMPUTEROWE WSPOMAGAJĄCE ANALIZĘ OBRAZÓW RTG W KONTROLI CELNO-SKARBOWEJ 95

SŁOWO WSTĘPNE

Ten numer Przeglądu Nauk Stosowanych poświęcony jest w całości ogólnopolskiej konferencji naukowej zatytułowanej „Praktyczne aspekty i możliwości wykorzystania potencjału naukowo- badawczego oraz transfer wiedzy pomiędzy sektorem nauki, a jednostkami Krajowej Administracji Skarbowej", która odbyła się w Opolu w dniach 13-14 marca 2018 r. z inicjatywy Izby Administracji Skarbowej w Opolu i Politechniki Opolskiej. W niniejszym numerze Przeglądu Nauk Stosowanych zamieszczono informacje o konferencji oraz opublikowano wybrane artykuły autorów wystąpień konferencyjnych. Wśród autorów artykułów są zarówno pracownicy naukowi uczelni, jak również pracownicy i funkcjonariusze jednostek Krajowej Administracji Skarbowej. Wiele artykułów posiada dwóch autorów reprezentujących obydwa środowiska, co dowodzi współpracy pomędzy tymi sektorami.

Głównym celem tego naukowego wydarzenia była dyskusja i wymiana doświadczeń pomiędzy środowiskiem naukowym uczelni a przedstawicielami izb administracji skarbowej z całego kraju dotycząca możliwych form i obszarów zacieśnienia współpracy obu środowisk. W trakcie dwóch dni konferencji teoretycy i praktycy mogli spotkać się i podyskutować o możliwościach oraz korzyściach, jakie daje partnerstwo nauki z administracją skarbową. Ku satysfakcji Organizatorów konferencja charakteryzowała się wysokim poziomem merytorycznym dyskusji, a jej tematyka spotkała się z dużym zainteresowaniem przedstawicieli obu środowisk. Mamy nadzieję, że publikacja będzie nie tylko źródłem wiedzy, dobrych praktyk,ale także inspiracją dla innych jednostek administracji publicznej.

Paweł Cybulski
Podsekretarz Stanu
Zastępca Szefa Krajowej Administracji Skarbowej
Ministerstwo Finansów
ul. Świętokrzyska 12
00-916 Warszawa
pawel.cybulski@mf.gov.pl

Przemysław KRAWCZYK
Przemysław MISIURSKI

ANALIZA DANYCH PODATKOWYCH – ZARYS PROBLEMU

Streszczenie: W artykule przedstawiono zarys problemu analizy danych podatkowych wynikający z przetwarzania przez organy administracji skarbowej coraz to większej liczby danych pochodzących m.in. z Jednolitych Plików Kontrolnych (JPK). Zaprezentowano teoretyczne aspekty eksploracji danych oraz przedstawiono dostępne oprogramowanie pozwalające na dokonywanie złożonych analiz. W dalszej kolejności przedstawiono narzędzia stosowane w pracy analityka administracji skarbowej.

Słowa kluczowe: eksploracja danych, analiza danych, JPK

ANALYSIS OF TAX DATA - OUTLINE OF THE PROBLEM

Summary: The article presents an outline of the problem of tax data analysis resulting from the processing by tax administration authorities of an ever-increasing number of data originating, i.a. from Standard Audit File for Tax (JPK). The theoretical aspects of data mining and the available software allowing for complex analyses are presented. Subsequently, the tools used in the work of the tax administration analyst are presented.

Keywords: data mining, data analysis, Standard Audit File for Tax

1. WSTĘP

Celem niniejszego artykułu jest przedstawienie podstawowych informacji nt. analizy danych podatkowych. Administracja skarbowa stara się wdrożyć zcentralizowany model zarządzania danymi. Model ten opiera się na założeniu, że tylko w ten sposób można zapewnić odpowiednią wydajność poprzez tworzenie dużych raportów (raporty oparte o dane z JPK), które po dodatkowej „obróbce" przekazywane są do użytkowników końcowych tj. pracowników działów czynności sprawdzających lub analiz odpowiednio urzędów skarbowych i urzędów kontroli celno-skarbowej.

Systemy informatyczne, w tym systemy analityczne, powinny mocniej wspierać administrację skarbową w realizacji celów. Celami tymi jest zapewnienie bezpieczeństwa finansowego Polski, w tym efektywnego poboru podatków i ceł. Realizacji tych celów służy proces centralizacji baz danych, centralizacji procesu analiz, monitorowania poziomu ryzyka podatkowego, a także centralizacja typowania podmiotów do kontroli. Obecnie ogromnym wyzwaniem jest dynamiczny wzrost wolumenu gromadzonych i przetwarzanych przez administrację skarbową danych dotyczących podatników, ich zachowań oraz szeroko

pojętego otoczenia biznesowego. Stąd od jakości zastosowanych narzędzi informatycznych w dużej mierze zależy efektywność działań.

Proces centralizacji baz danych opiera się na stworzeniu jednego źródła danych w postaci tzw. „Fundamentu Danych", na którym gromadzone byłyby dane wykorzystywane w analizach. Proces ten obejmuje konieczność integracji z kilkudziesięcioma źródłami danych (baz danych). Rozwiązanie to obejmuje dane zarówno o podatnikach, deklaracje podatkowe, ewidencje JPK-VAT składane co miesiąc i zawierające zestawianie wszystkich faktur oraz inne informacje gromadzone lub wytwarzane przez administrację skarbową.

Fundament Danych pozwoli nie tylko na zgromadzenie danych w jednym miejscu, ale także umożliwi optymalizację wykonywania obliczeń. W powyższym celu wykorzystywane są rozwiązania analityczne, oparte na relacyjnych bazach danych, jak i na technologii baz grafowych. Możliwe jest stosowanie szerokiego spektrum technik analitycznych, w szczególności technik eksploracji danych oraz technik SNA analizy sieciowej lub społecznej analizy sieciowej badania sieci społecznej, wykorzystujące teorię sieci i koncentrujące się na analizie stosunków pomiędzy elementami sieci (jednostkami, organizacjami itp.). W analizie sieciowej nacisk kładzie się na relacje i ich wzorce, z których wynikają szanse i ograniczenia dla węzłów sieci.

2. JEDNOLITY PLIK KONTROLNY - ZAKRES DANYCH WEJŚCIOWYCH

Zgodnie z wytycznymi międzynarodowych organizacji gospodarczych takich jak OECD[11], IMF[12] w ostatnich latach rozpoczął się proces upraszczania i ujednolicania rozliczeń podatkowych. Działania te mają na celu doprowadzić do poprawy przejrzystości transakcji finansowych. W związku z tymi działaniami, jednym z rozwiązań zarekomendowanych przez OECD, jest jednolity, międzynarodowy standard rozliczeń podatkowych, tzw. SAF-T (Standard Audit File for Tax). W wielu krajach europejskich funkcjonuje już SAF-T. Również Polska w 2016 roku wprowadziła swój model raportowania danych tzw. Jednolity Plik Kontrolny. Dzięki niemu organy administracji skarbowej mogą w łatwy sposób analizować sytuację podatkową firm i upłynnić ściągalność podatku VAT [ttps://businessinsider.com.pl/firmy/przepisy/ile-firm-zlozylo-jpk-vat-za-styczen-2018/01htqrn]. Jednolity Plik Kontrolny jest zbiorem (bazą) danych, tworzonym z systemów informatycznych przedsiębiorcy, zawierającym informacje o operacjach gospodarczych za dany okres, mającym układ i format umożliwiający jego łatwe przetwarzanie. Inaczej mówiąc, jest to standard, według którego przekazuje się dane do organów podatkowych [http://chem-eng.utoronto.ca/~datamining/dmc/data_mining.htm].

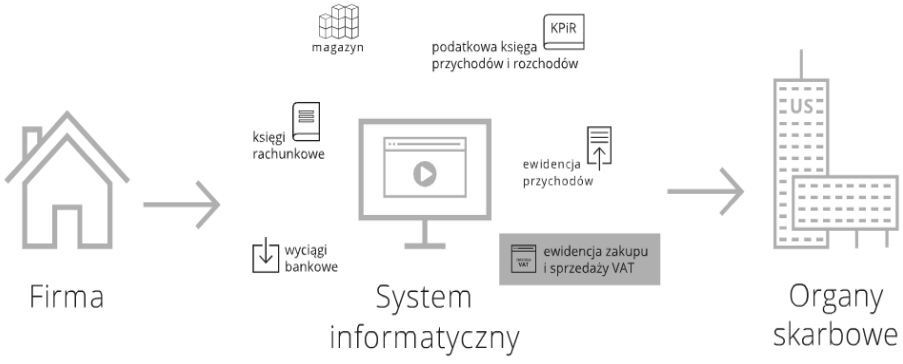
[11] *Organisation for Economic Co-operation and Development - OECD - Organizacja Współpracy Gospodarczej i Rozwoju*

[12] International Monetary Fund - IMF Międzynarodowy Fundusz Walutowy

Jednolity Plik Kontrolny został wprowadzony ustawą z 10 września 2015 r. o zmianie ustawy – Ordynacja Podatkowa (Dz.U. z 2015 r. poz. 1649) w dodanym art. 193a, który w *§ 1* mówi: w przypadku prowadzenia ksiąg podatkowych przy użyciu programów komputerowych, organ podatkowy może żądać przekazania całości lub części tych ksiąg oraz dowodów księgowych za pomocą środków komunikacji elektronicznej lub na informatycznych nośnikach danych, w postaci elektronicznej odpowiadającej strukturze logicznej, o której mowa w § 2, wskazując rodzaj ksiąg podatkowych oraz okres, którego dotyczą [Ustawa z dnia 10 września 2015r.].

Elektroniczne raportowania miesięczne na cele podatku VAT (bez wezwania organu podatkowego) są obowiązkowe dla dużych jednostek od lipca 2016, małe i średnie podmioty obowiązek takiego raportowania mają od stycznia 2017 [Voss G. 2017]. Od 1 stycznia 2018 roku wszyscy przedsiębiorcy zarejestrowani jako czynni podatnicy podatku VAT (w tym mikroprzedsiębiorcy), mają obowiązek przesłania do organów skarbowych tzw. JPK_VAT w formie elektronicznej (poglądowy schemat przekazania niezbędnych danych prezentuje rysunek 1). Według Ministerstwa Finansów za miesiąc styczeń pliki JPK_VAT przesłało 1,5 mln przedsiębiorców [ttps://businessinsider.com.pl/firmy/przepisy/ile-firm-zlozylo-jpk-vat-za-styczen-2018/01htqrn].

Rysunek 1. Schemat przekazania danych JPK



Źródło: http://www.edat.pl/enova365/jednolity-plik-kontrolny

Formalności związane z Jednolitym Plikiem Kontrolnym wymusiły na przedsiębiorcach stosowanie odpowiednich systemów informatycznych, które ułatwiają przesłanie danych do organów administracji skarbowej. Tak duża ilość danych wysyłana przez podmioty gospodarcze wymaga również od instytucji skarbowych stosowania odpowiednich narzędzi informatycznych przeznaczonych do ich analizy i ich eksploracji w celu pozyskania odpowiedniej wiedzy analitycznej.

3. EKSPLORACJA DANYCH

Obecnie jedną z najdynamiczniej rozwijanych dziedzin informatyki jest eksploracja danych (ang. data mining). Duże zainteresowanie eksploracją danych wynika w głównej mierze z problemu efektywnego i racjonalnego wykorzystania danych nagromadzonych w bazach danych przez ogromną liczbę przedsiębiorstw, jak również instytucji administracji publicznej czy ośrodków naukowych [Morzy T. 1999].

Według definicji eksploracja danych to proces odkrywania wzorców, reguł, zależności w dużych zbiorach danych (hurtownie danych). Zasadniczym celem eksploracji danych jest wydobycie nowej – nieznanej informacji z baz danych lub ze zbiorów wiedzy [Olszak C.M., Bartuś K. 2009], [Racka K. 2015]. Eksploracja danych integruje wiele dyscyplin wśród których można wymienić: statystykę, sztuczną inteligencję, systemy bazodanowe, optymalizację [Morzy T. 1999].

Rysunek 2. Eksploracja danych

Źródło: opracowanie własne n/p http://chem-eng.utoronto.ca/~datamining/dmc/ data_mining.htm.

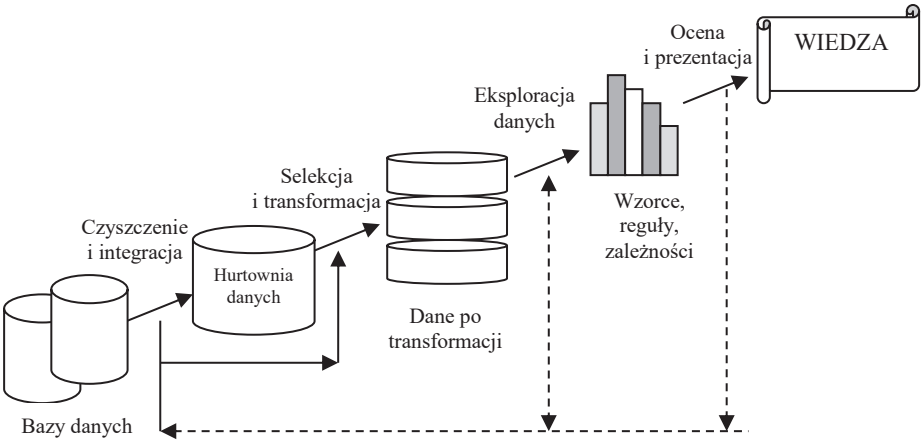
Eksploracja danych stanowi jeden z etapów procesu tworzenia wiedzy z baz danych. Proces ten składa się z następujących kroków [Morzy T. 1999], [Racka K. 2015]

- czyszczenie danych - usuwanie błędnych danych wynikających z pomyłek operatora lub danych powielonych;
- integracja danych - łączenie danych pochodzących z różnych źródeł o niejednolitej strukturze i różnorakich modelach danych;
- wybieranie danych - wybór danych niezbędnych do przeprowadzenia analizy;

- transformacja danych - przetworzenie lub łączenie danych w formy odpowiednie do eksploracji;
- eksploracja danych - wydobycie z danych odpowiednich wzorców i zależności
- ocena i prezentacja odkrytych wzorców, reguł, zależności - identyfikacja odkrytych wzorców i prezentacja odkrytej wiedzy docelowym użytkownikom.

Zależności pomiędzy eksploracją danych a pozostałymi etapami procesu pozyskiwania wiedzy prezentuje rysunek 3.

Rysunek 3. Eksploracja danych jako jeden z kroków w procesie odkrywania wiedzy



Źródło: Opracowanie własne na podstawie [Ejsmont K., Krystosiak K., Lipiak J. 2015], [Han J., Kamber M. 2001], [Racka K. 2015].

Eksploracja danych jest najistotniejszym elementem procesu pozyskiwania wiedzy. Pozostałe etapy tego procesu są albo czynnościami rutynowymi w porównaniu do eksploracji danych, albo są nierozłącznie związane z samą eksploracją [Ejsmont K., Krystosiak K., Lipiak J. 2015]. W literaturze przedmiotu wymienia się sześć metod eksploracji danych [Morzy T. 1999], [Racka K. 2015]:

1. **Wyszukiwanie asocjacji -** jest to najszersza klasa metod obejmująca odkrywanie różnego rodzaju nieznanych zależności w bazie danych. Metody w tej klasie obejmują głównie odkrywanie asocjacji pomiędzy obiektami.
2. **Klastrowanie - metoda grupowania** - celem tych metod jest znajdowanie w bazie danych skończonych podzbiorów (klas, grup) , które posiadają podobne cechy. W metodach tych liczba potencjalnych klastrów nie jest znana, stąd, proces grupowania przebiega, najczęściej, w dwóch cyklach: cykl zewnętrzny przebiega po liczbie możliwych klastrów, cykl wewnętrzny próbuje znaleźć optymalny podział obiektów pomiędzy klastry.

3. **Odkrywanie wzorców sekwencji** - celem tych metod jest odkrywanie czasowych wzorców zachowań, na podstawie analizy danych zmieniających się w czasie.
4. **Odkrywanie klasyfikacji** - celem tych metod jest znajdowanie zależności pomiędzy klasyfikacją obiektów a ich charakterystyką.
5. **Odkrywanie podobieństw w przebiegach czasowych -** celem tych metod jest znajdowanie podobieństw w przebiegach czasowych opisujących określone procesy.
6. **Wykrywanie zmian i odchyleń -** metody te pozwalają na znajdowanie różnic pomiędzy aktualnymi a oczekiwanymi wartościami danych.

W praktyce stosowanie różnych metod eksploracji danych dla tego samego zagadnienia może okazać się korzystniejsze, gdyż zastosowanie jednej metody może nie być wystarczające do całościowego rozwiązania rozpatrywanego problemu [Racka K. 2015].

Do skutecznego przeprowadzenia procesu analizy danych, prócz dobrej znajomości badanej dziedziny, zbioru danych, czy wybrania właściwej techniki eksploracji, niezbędnym jest wykorzystanie odpowiedniego oprogramowania.

Na rynku oprócz programów komercyjnych takich firm jak: HP, IBM, MICROSFT, ORACLE, SAS Institute, StatSoft dostępna jest duża liczba programów niekomercyjnych oferujących rozwiązania z zakresu eksploracji danych [Racka K. 2015]. Listę przykładowych programów do eksploracji danych przedstawia tabela 1.

Tabela 1. Lista przykładowych darmowych programów do eksploracji danych

Nazwa programu	Typ licencji
CMSR DATA Miner	Licencja akademicka na trzy lata, wersja darmowa na 6 miesięcy
Databionic ESOM Tools	GNU GPL
ELKI	AGPL
KNIME	GNU GPL
Mloss	GNU GPL
Mlpy	GNU GPL
Orange	GNU GPL
Projekt R	GNU GPL
Rapid Miner	AGPL/ Proprietary (prawnie zastrzeżone)
Rattle GUI	GNU GPL
SCaViS	jądro silnika programu GPL, instalacja, dokumentacja, podzespoły darmowe ale nie do celów komercyjnych
SenticNEt API	Darmowy z umieszczeniem informacji: Copyright © 2012 Yuri Malheiros
Weka 3	GNU GPL

Źródło: [Racka K. 2015]

Duża grupa programów wymienionych w powyższej tabeli to programy na licencji GNU GPL, która daje możliwość użytkownikowi uruchamiania programu dowolną liczbę razy, a udostępniony kod źródłowy programu można modyfikować na własne potrzeby, a także rozpowszechniać. Natomiast licencja AGPL jest licencją wolnego oprogramowania uruchamianego przez sieć. Programy na licencji AGPL są rozbudowywane i zmieniane na potrzeby odbiorców przez dużą liczbę informatyków [Racka K. 2015].

Do głównych zalet niekomercyjnych programów zalicza się przede wszystkim fakt, że użytkownicy mają do nich darmowy dostęp, a praca na tych programach daje wiele możliwości obliczeniowych użytkownikom dorównując, a często przewyższając jakościowo programy komercyjne [Racka K. 2015].

4. ANALIZA DANYCH PODATKOWYCH - WYKORZYSTYWANE NARZĘDZIA

Wykorzystywane w pracy analityka administracji skarbowej narzędzia, to oprócz programów komercyjnych, także KNIME i Neo4j. KNIME Analytics Platform jest otwartym oprogramowaniem analitycznym wykorzystywanym w celu integracji rozwiązań i technologii informatycznych oraz informacyjnych. Pozwala na wykonywanie analiz i obliczeń statystycznych, jak również drążenie danych, wykrywanie wiedzy z danych, stosowanie metod sztucznej inteligencji, wdrażanie automatycznych rozwiązań analitycznych czy procesów ETL. Platforma udostępniona jest na licencji GNU General Public License, Version 3. Posiada ponad 2000 modułów, setki gotowych do uruchomienia przykładów, możliwość zwiększenia funkcjonalności poprzez instalacje dodatkowych rozszerzeń.

KNIME jest ciekawym rozwiązaniem dla badaczy danych. Posiada forum dyskusyjne, kanał na YouTube oraz pełną dokumentację dostępną z poziomu aplikacji. Pakiety instalacyjne dostępne są dla systemów: Microsoft Windows, Linux, Mac. Istnieje również wersja SDK pozwalająca tworzyć własne moduły.

KNIME jest oprogramowaniem opartym na graficznym interfejsie użytkownika. Procesy tworzy się głównie w sposób graficzny poprzez wykorzystanie gotowych modułów (nodes), które połączone ze sobą tworzą przepływy analityczne (Workflow). Takie podejście pozwala na łatwe zrozumienie zadań wykonywanych w przepływie, powtarzalność analiz, zapewnienie jakości danych oraz pewność wyników.

Integracja rozwiązań zapewnia ogromną funkcjonalność narzędzia. W Knime korzystając z rozszerzeń można m.in. wykonywać skrypty pisane w językach: java, python, R czy perl. Dodatki umożliwiają dostęp do algorytmów z aplikacji WEKA, wszelkich relacyjnych baz danych, takich jak: Microsoft SQL Server, IBM DB2, Oracle DB, PostgreSQL, MySQL, Firebird, SQLite oraz innych posiadających sterownik JDBC. Można również uzyskać dostęp do systemów Big Data jak Apache Hadoop, Hive, grafowych baz danych (Neo4j) czy baz NoSQL (MongoDB). Rozszerzenia powodują, że jest to kompletne środowisko dla analityka.

KNIME pozwala na wykorzystanie metodologii takich jak: analizy statystyczne, big data, text mining, patterns matching, web mining, ETL (Extract, Transform and Load), EDA - exploratory data analysis), Data mining, SNA – (Social network analysis), GIS – wizualizację danych na mapach, wizualizację danych (wykresy) oraz tworzenie raportów opartych na przygotowanych szablonach.

Neo4j to grafowa baza danych umożliwiająca przechowywanie grafów, pozwalająca na bardzo szybkie i proste przeszukiwanie zależności oraz powiązań. Platforma posiada wersję „Neo4j Community Edition" opartą na licencji GNU General Public License, Version 3.

Zaletą jest brak sztywnych tabel czy schematów znanych z relacyjnych baz danych. Baza posiada 4 podstawowe typy obiektów: NODE, RELATIONSHIPS, PROPERTIES, LABELS. Nodes reprezentują dowolne obiekty umieszczone w bazie, a RELATIONSHIPS określają relacje/zależności pomiędzy obiektami. Obiekty oraz relacje mogą posiadać dodatkowe atrybuty (PROPERTIES). Szczególnym typem jest LABELS – semantyczny typ danych dla obiektów i relacji.

Praca na bazie wykorzystuje język CQL (Cypher Query Language) podobny do SQL, prosty w nauce i wykorzystaniu języka zapytań i manipulacji na danych. Nawiązuje do ascii-art. Interfejs dostępowy do bazy jest możliwy poprzez CLI (linię komend), przeglądarkę – własny portal www lub interfejs API.

Baza pozwala na dużo szybsze wyszukiwanie skomplikowanych powiązań bazując na grafach, zastępując skomplikowane, wielopoziomowe złączenia JOIN znane z SQL i relacyjnych baz danych.

Język Cypher pozwala na dopasowywanie obiektów i relacji. Pozwala tworzyć, uaktualniać, usuwać obiekty, relacje, nazwy semantyczne i właściwości. Neo4j posiada również możliwość wykorzystania wyzwalaczy (constraints) i indeksów w celu zwiększenia wydajności. Na danych można wykonywać operacje arytmetyczne czy agregację danych.

Bazę można zasilać wykorzystując język Cypher, importować dane z plików CSV oraz baz danych.

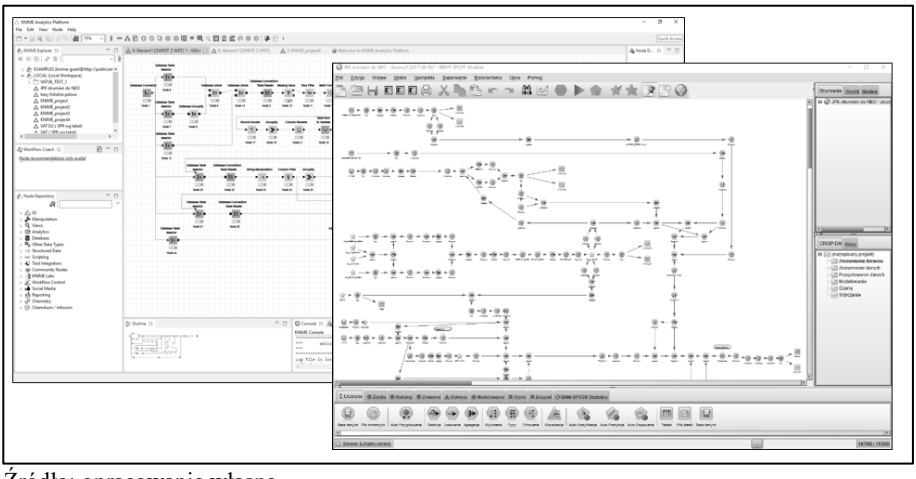
Baza wspiera również możliwość instalacji rozszerzeń zwiększających funkcjonalność. Daje to np. bezpośredni dostęp do danych z baz relacyjnych, dodatkowych algorytmów analiz sieciowych, interfejsów dla języków programowania, takich jak java, python, dotnet, ruby czy php. Możliwa jest wizualizacja danych oraz analiza w aplikacjach dedykowanych do SNA jak Gephi.

5. PROCES ANALIZY DANYCH PODATKOWYCH

Celem optymalnego wykorzystania stosowanych technik analitycznych jest dwuetapowe, hybrydowe podejście, łączące najlepsze praktyki w zakresie identyfikacji ryzyka na poziomie obiektów oraz przepływów. Wyraźne rozdzielenie warstwy obiektów oraz przepływów gwarantuje skalowalność oraz pełną konfigurowalność na podstawie przyjętych warunków brzegowych.

Identyfikacja podmiotów podwyższonego ryzyka oraz łańcuchów/sieci powiązań transakcyjnych z wykorzystaniem wyżej opisanych rozwiązań jest procesem złożonym. W pierwszym kroku prowadzona jest integracja oraz podstawowe czyszczenie danych, pozwalające na wygenerowanie zbioru analitycznego. Kolejny etap polega na nałożeniu na warstwę obiektów (także warstwę powiązań) wyników reguł biznesowych – indukowanych i ewaluowanych m.in. z wykorzystaniem algorytmów drzew decyzyjnych, pozwalających na oznaczenie podmiotów (stanowiących wierzchołki sieci transakcyjnej), spełniających cechy wskazujące na ryzyko udziału w procederze wyłudzenia skarbowego wraz ze wskazaniem prawdopodobnej roli obiektu.

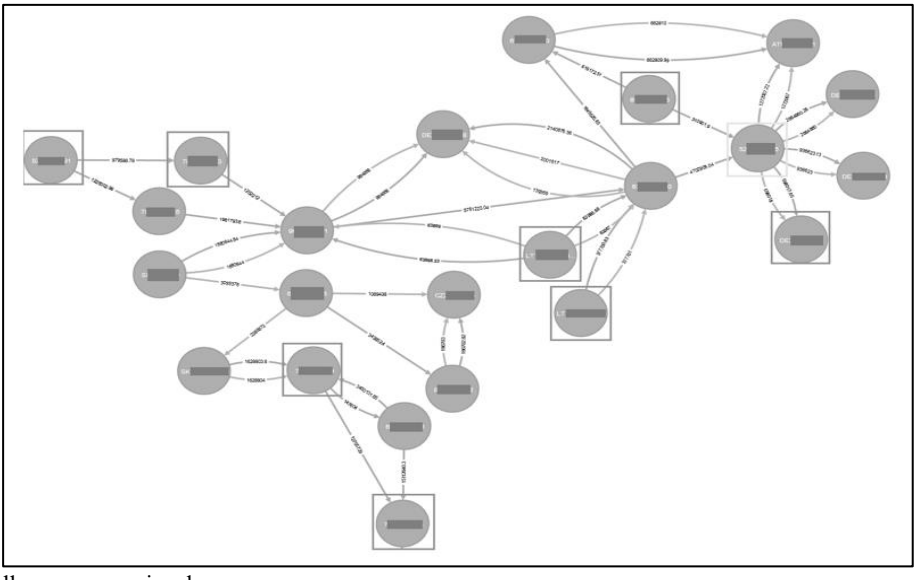
Rysunek 3. Przykład zastosowania narzędzi do analizy danych



Źródło: opracowanie własne.

W drugim etapie następuje integracja ww. zbiorów w bazie grafowej Neo4j, a także wygenerowanie sieci, zawierających obiekty, spełniające wyspecyfikowane we wcześniejszych krokach kryteria reguł biznesowych. Dla tak przedstawionych relacji generowane jest w narzędziu analitycznym otoczenie dalsze, pozwalające zidentyfikować potencjalne źródła towaru lub beneficjentów oszustwa.

Rysunek 4. Przykład zastosowania bazy grafowej Neo4j



Źródło: opracowanie własne

Docelowo w warstwie analitycznej rozwiązanie takie gwarantuje możliwość wykorzystania nowoczesnych technik statystycznej analizy danych wykraczających poza elementarne operacje matematyczne oraz standaryzację formy uzyskiwanych wyników wpływającą bezpośrednio na efektywność oceny operacyjnej podmiotu kwalifikowanego/podatnika lub zorganizowanej grupy podmiotów działających w celu wykorzystania sektora finansowego do wyłudzeń skarbowych

6. PODSUMOWANIE

Rozważania zaprezentowane w niniejszym artykule potwierdzają, że metody eksploracji danych mają zastosowanie w wielu obszarach nauki i życia gospodarczego. Są również wykorzystywane przez jednostki administracji skarbowej w celu efektywniejszej analizy danych skarbowych wysyłanych do urzędów przez podmioty gospodarcze. Przedstawiony w artykule opis stanowi jedynie zarys zagadnienia analizy danych podatkowych. Ciągle wyzwaniem jest zapewnienie dobrej jakości danych, w tym kwestia czyszczenia danych. Praca z danymi podatkowymi stanowi ogromne wyzwanie. Aby sprostać temu wyzwaniu pożądane jest zapewnienie odpowiedniej współpracy z ośrodkami akademickimi.

Literatura:

[1] Ejsmont K., Krystosiak K., Lipiak J.: *Zastosowanie wybranej techniki eksploracji danych w przemyśle poligraficznym*, Opole, Innowacje w Zarządzaniu i Inżynierii Produkcji. T.2*.,* 2015.

[2] Han J., Kamber M.: *Data mining: Concepts and Techniques*, Morgan Kaufmann Publishers, Academic Press, 2001.

[3] Morzy T.: *Eksploracja danych: problemy i rozwiązania,* Zakopane, V Konferencja PLOUG, 1999.

[4] Olszak C.M., Bartuś K.: *Analiza i ocena wybranych modeli eksploracji danych*, Opole, Komputerowo Zintegrowane Zarządzanie. Tom II. 2009.

[5] Racka K.: *Metody eksploracji danych i ich zastosowanie*, Zeszyty Naukowe PWSZ w Płocku, Nauki Ekonomiczne, t. XXI, 2015.

[6] Świder K., Jędrzejec B.: *Zaawansowane metody analizy danych i niekomercyjne pakiety analityczne w systemach wspomagania decyzji na potrzeby administracji publicznej źródła internetowe*, Warszawa, Technologie informatyczne w administracji publicznej, KAE SGH, 2014.

[7] Ustawa z dnia 10 września 2015 r. *o zmianie ustawy – Ordynacja Podatkowa* (Dz.U. z 2015 r. poz. 1649 z poźn. zm.).

[8] Voss G.: *Rachunkowość w procesie cyfryzacji - obszary ryzyka*, Warszawa, Studia i prace Kolegium Zarządzania Finansów Zeszyt Naukowy 157, 2017.

Źródła internetowe

[9] Business Insider Polska ttps://businessinsider.com.pl/firmy/przepisy/ile-firm-zlozylo-jpk-vat-za-styczen-2018/01htqrn [dostęp 18.06.2017].

[10] Data mining http://chem-eng.utoronto.ca/~datamining/dmc/data_mining.htm [dostęp 27.06.2017]

[11] Edat.pl http://www.edat.pl/enova365/jednolity-plik-kontrolny [dostęp 26.06.2017]

[12] Kariera w finasach, *Big data w służbie fiskusa: czas na globalny urząd skarbowy?* https://www.karierawfinansach.pl/artykul/wiadomosci/big-data-w-sluzbie-fiskusa-czas-na-globalny-urzad-skarbowy [dostęp 16.06.2017].

[13] Serwis informacyjno-usługowy dla przedsiębiorców Biznes.gov.pl, https://www.biznes.gov.pl/pl/firma/podatki-i-ksiegowosc/chce-prowadzic-ksiegowosc/jednolity-plik-kontrolny-jpk [dostęp 16.06.2017].

Przemysław Krawczyk
Dyrektor Departamentu Nadzoru nad Kontrolami
Krajowa Administracja Skarbowa

dr inż. Przemysław Misiurski
Politechnika Opolska
Wydział Ekonomii i Zarządzania
ul. Luboszycka 7, 45-036 Opole
e-mail: p.misiurski@po.opole.pl

POLITECHNIKA OPOLSKA

ISSN 2353-8899