Gliwice, dn.12.05.2026 r.

Prof. dr hab. inż. Marcin Gorawski
Katedra Informatyki Stosowanej
Wydział Automatyki, Elektroniki i Informatyki
Politechnika Śląska
ul. Akademicka 16,
44-100 Gliwice,
Marcin.Gorawski@polsl.pl

Recenzja rozprawy doktorskiej
mgr Macieja Brzeskiego
zatytułowanej:
Analiza podobieństwa procesów ETL i pochodzenia danych w kontekście zarządzania hurtownią danych

Podstawą opracowania recenzji jest pismo Przewodniczącego Rady Dyscypliny Informatyki Technicznej i Telekomunikacji Uniwersytetu Jagiellońskiego (RDITT UJ) nr 1199.5101.35.2026 z dnia 29.01.2026 r. w związku, i wraz z uchwałą RDITT UJ (nr 1a/01/2026) z dnia 22.01.2026 r. W uchwale tej powołano mnie na recenzenta w postępowaniu o nadanie stopnia doktora panu magistrowi Maciejowi Brzeskiemu. Rozprawa doktorska (doktorat wdrożeniowy) liczy sobie 212 stron i jest przygotowana w języku polskim. Uzyskałem wgląd do papierowej wersji rozprawy i oryginalnego pisma RDITT UJ w dniu 12.02.2026 r. (pliki pdf a formalnych dokumentów (trzy egzemplarze umowy, rachunek, oświadczenie RODO, oświadczenie dot. instrukcji i oświadczenie dla celów podatkowych) w dniu 25.02.2026r. Podpisanie i odesłanie umowy w wersji papierowej pocztą (list polecony) nastąpiło w dniu 2.03.2026 r. Zatem ustawowy dwumiesięczny termin sporządzenia recenzji mija w dniu 2.05.2026 r.

1. **Dane o doktorancie:**

A. Studia

Absolwent studiów licencjackich w Uniwersytecie Jagiellońskim, na kierunku:
i. matematyka teoretyczna, w latach od 2008 r. do 2012 r.,
ii. informatyka analityczna, w latach od 2012 r. do 2015 r.

Absolwent studiów magisterskich na kierunku matematyka finansowa w Uniwersytecie Jagiellońskim, w latach od 2012 r. do 2013 r.

W latach 2020–2025 r. odbył studia doktoranckie w Szkole Doktorskiej Nauk Ścisłych i Przyrodniczych dla programów kształcenia: informatyka, matematyka Wydziału Matematyki i Informatyki UJ.

B. Rodzaj doktoratu

Przedstawiona do recenzji rozprawa doktorska mgr Macieja Brzeskiego pt.: *Analiza podobieństwa procesów ETL i pochodzenia danych w kontekście zarządzania hurtownią danych* dokumentuje przebieg realizacji doktoratu wdrożeniowego. Rozprawa ma charakter wdrożeniowy, przygotowana w ramach programu "Doktorat wdrożeniowy" o symbolu projekt DWD/4/66/2020, finansowanego przez Ministerstwo Nauki i Szkolnictwa Wyższego. Ten doktorat wdrożeniowy był realizowany we współpracy z firmą Informatica – polska firma (to nie Informatica Polska – polski dystrybutor Informatica Corp powstała w Stanach Zjednoczonych w 1993 roku). Motywacją podjęcia tej tematyki rozprawy jest obszar strategii projektowania i zarządzania procesami ETL wielkich skali (zamiennie: wieloskalowe procesy ETL)

(ang. *large-scale ETL processes, lsETL, ETL(sl), ETL^{sl}*) dla dużych organizacji, w celu automatyzacji i optymalizacji kosztów uzyskania informacji o pewności wiarygodnego pochodzenia i przepływu danych. Doktorant sformułował (nie wprost) tezę: Możliwe jest poprzez specjalizowaną analizę mechanizmów podobieństwa procesów ETL oraz użycia metod rekonstrukcji i uzupełnianie brakujących danych zwiększyć jakość wiedzy o danych w hurtowniach danych, co przekłada się zarówno na ich praktyczne utrzymanie, jak i na zaufanie do wyników analiz biznesowych.

C. Promotor i opiekun pomocniczy

Rozprawa wykonana została pod kierunkiem promotora prof. dr hab. Adama Romana, kierownika Katedry Inżynierii Oprogramowania w Instytucie Informatyki i Matematyki Komputerowej Wydziału Matematyki i Informatyki UJ oraz opiekuna pomocniczego mgr Dawida Dudy.

D. Doświadczenie zawodowe - miejsce pracy z podaniem zajmowanych stanowisk lub pełnionych funkcji

Mgr Maciej Brzeski wykazał aktywność zawodową i akademicką. Mgr Maciej Brzeski pracuje zawodowo od 10 lat jako inżynier oprogramowania (SE), w tym na stanowisku: programisty (1), starszego (1,5) i głównego (5) inżyniera oraz kierownika zespołu inżynierów oprogramowania (5)

Aktywność zawodowa

Bezpośrednia praktyka w tematyce recenzowanej rozprawy mgr Maciej Brzeski to 6 letnia praca w firmie Informatica Sp. z o.o. (wcześniej *Compact Solutions Poland sp z o.o.)* na stanowisku:

a. **Główny inżynier oprogramowania** odpowiedzialny za wdrażanie strategii techniczno-badawczej zorientowanej na ML (ang. Machine Learning, ML) oraz automatyzację mechanizmów pochodzenia danych (ang. Data Lineage, DL) w systemach zarządzanie danymi z użyciem AI.
b. **Kierownik zespołu inżynierów oprogramowania** odpowiedzialny za prace zarówno nad automatycznym wnioskowaniem o pochodzeniu danych i modelami NLP opartych na architekturze Transformer, jak i użyciem LLM do precyzyjnej ekstrakcji metadanych z dokumentów i arkuszy kalkulacyjnych.
c. **Starszy inżynier oprogramowania** d/s opracowania metod opartych na LM do analizy ETL^{sl}, w tym: a) Identyfikacja grup podobnych procesów integracji danych dla parametryzowanych i wielokrotnego użytku przepływów pracy, b) porównywanie złożonych procesów z użyciem zmodyfikowanych odległości edycyjnych drzew i grafów, c) Projektowanie skalowalnych metod wizualizacji bardzo dużych grafów procesów (do 100 000 węzłów).

Pośrednia praktyka mgr Maciej Brzeski to dwuletnia praca w firmie Crif Sp. z o.o. oraz Teroplan S.A. jako programista systemów backendowych i systemów bazodanowych klasy enterprise oraz systemów przetwarzania i automatyzacji przepływów danych transportowych.

Aktywność akademicka

Mgr Maciej Brzeski jako specjalista/badacz ML uczestniczył 4 lata w projekcie badawczo-rozwojowym TensorCell (2018-2021, w trybie zdalnym) z zakresu optymalizacji ruchu drogowego oraz logistyki. Doktorant prowadził zajęcia na Uniwersytecie Jagiellońskim w latach 2015–2023 r.

z przedmiotów: a) Uczenie maszynowe, b) Algorytmy i struktury danych, c) Rachunek prawdopodobieństwa i statystyka.

2 **Cel rozprawy**

Celem pracy doktorskiej mgr Macieja Brzeskiego pt.: ***Analiza podobieństwa procesów ETL i pochodzenia danych w kontekście zarządzania hurtownią danych*** było podjęcie próby podniesienia jakości i niekompletności informacji o przepływach danych w procesach ETL wielkiej skali.

Obecnie w przedsiębiorstwach liczba procesów ETL realizowana przez silniki ekstrakcji danych (silniki ETL) klasycznych systemów hurtowni danych (systemy DW) mieści się w zakresie 300 – 3000+ procesów ETL przy 30–200 systemach źródłowych (min. ERP, CRM, platforma bankowa, systemy produkcyjne, pliki, API, itd.) przy częstotliwości ich uruchamiania zwykle dziennym (podstawowe raporty), godzinnym (personalizacja, monitoring). Dla lepszego zrozumienia całościowego zagadnienia należy przestawić Systemy Wspomagania Podejmowania Decyzji (systemy DSS) bazujące na kluczowych architekturach systemów DW; niestety brak tej wiedzy w rozprawie jest powodem występowania poważnych niejednoznaczności używanych pojęć i definicji.

Ocena celu i tezy rozprawy

Trudno dokonać oceny celu i tezy rozprawy ponieważ opis koncepcji hurtowni danych (rozdz. 2.1.2) oraz procesy ETL i ELT (rozdz. 2.1.3) daje informacje popularno-naukowe. Skutkiem tego jest techniczne podejście do zagadnień badawczych w dziedzinie systemów DW. Z drugiej strony, w rozprawie pojawia się tylko 2 razy sformułowanie ***zarządzania hurtownią danych*** a mianowicie, w tytule rozprawy i na stronie 17 w cytowanym niezrozumiałym poniżej zdaniu.

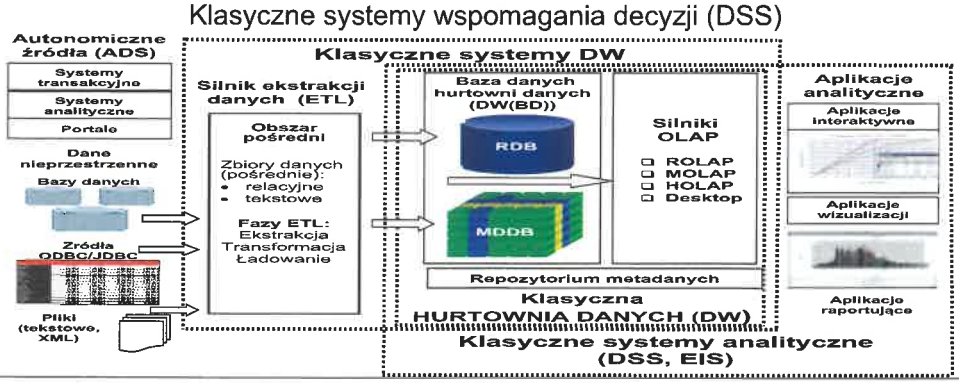
Cyt. S17: „W rezultacie tradycyjne podejście do ***zarządzania hurtownią danych*** okazuje się niewystarczające. Konieczne staje się wprowadzenie mechanizmów zapewniających przejrzystość, kontrolę i możliwość śledzenia danych w całym cyklu ich życia – co będzie przedmiotem dalszych rozdziałów tej pracy."

Zatem problemem oceny celu i tezy tej rozprawy staje się pojęciowe jego usadowienie w obszarze badań podstawowych i zaawansowanych przy uwzględnieniu kontekstów utylitarno-eksperymentalnych zorientowanych na procesy ETL. Stąd konieczne stało się przedstawienie poniższej ramowej wiedzy naukowo-badawczej systemów DW.

BADANIA PODSTAWOWE

Dla zrozumienia zagadnienia należy przestawić podstawowe architektury systemów DSS bazujące na hurtowniach danych (DW). Ramowy opis w wykładzie pod linkiem Wykład Profesorski - prof. dr hab. inż Marcin Gorawski. Wyróżniamy klasyczne i zaawansowane systemy DSS (EIS).

Klasyczne systemy wspomagania decyzji (DSS)



Klasyczne hurtownie danych (DW)

Autonomiczne źródła danych: **dane nieprzestrzenne** (zbiory – płaskie, relacyjne, tekstowe)
Model konceptualny: **jednotematyczny, fakty, wymiary, hierarchie atrybutów**
Model logiczny:

- **relacyjny OLAP** (ROLAP) (gwiazda, płatek śniegu, konstelacja faktów),
- **wielowymiarowy MOLAP** (MOLAP)

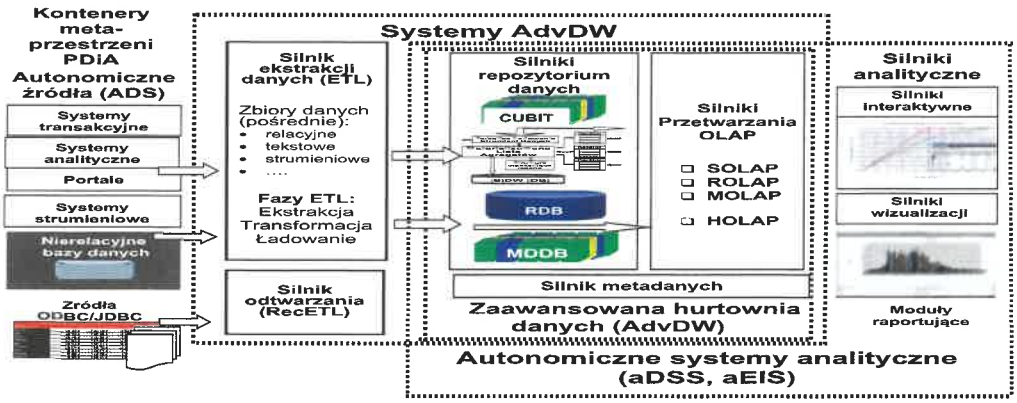
Model fizyczny:

- indeksy: ***bitmapowy, połączeniowy, bitmapowy połączeniowy***
- perspektywy zmaterializowane: ***metody odświeżania, zapytanie do tabel bazowych***
- partycjonowanie danych: ***round – robin, haszowe, wartościujace, hybrydowe***
- równoległe operacje: ***sortowanie danych, operacje I/O, polecenia SQL***

Klasyczny silnik ekstrakcji danych ETL (klasyczny silnik ETL) rozbudowany o silnik odtwarzania danych (silnik RecETL).

BADANIA ZAAWANSOWANE

Systemy przetwarzania wielkiej skali (systemy Big Data) wymagają zaawansowanych architektur systemów DSS bazujących na zaawansowanych hurtowniach danych (**systemy AdvDW**). Wyzwania badawcze-eksperymentalne i silne podstawy teoretyczne systemów AdvDW przedstawiono w nowej dziedzinie o nazwie **„Przestrzenie Danych i Algorytmów" (PDiA)**, gdzie zaprojektowano je jako Specjalizowane Architektury PDiA.



Systemy AdvDW zawierają komponenty:

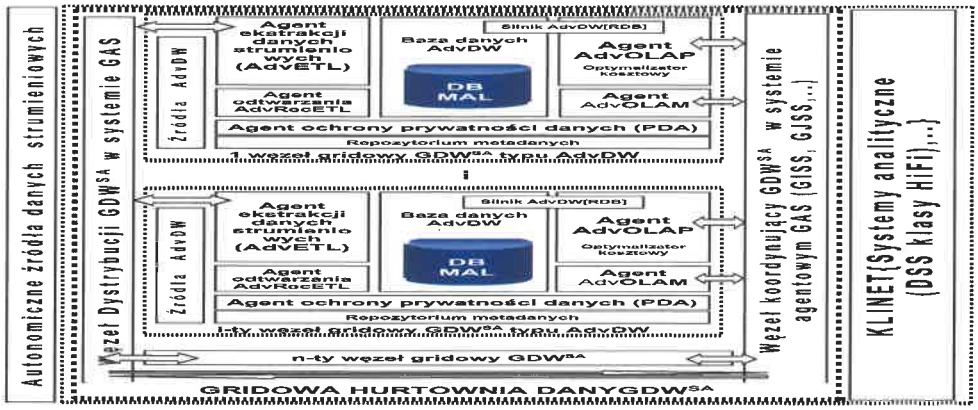
- Silniki kluczowe (ETL, StrETL, Rep. danych, OLAP, StrDW(MAL), CUBIT,....)
- Silniki atrybutowe (RecETL, StrRecETL, Rep. metadanych, SOLAM,...)
- Silniki ochrony prywatności danych (DPP, DPPMQ,....)
- Silniki kontekstowe (środowiskowe, przestrzenne, czasowe,...)
- Silniki anomalne (krytyczne, pozaklastrowe,....)
- Silniki metadanych
- Kontenery metaprzestrzeni PDiA

Wyróżniamy kilka klas systemów AdvDW o różnym stopniu skomplikowania architektury optymalnego przetwarzania analitycznego, hybrydowych zbiorów danych w tym danych strumieniowych i przestrzennych. W każdej z 4 klas AdvDW występują zaawansowane silniki ekstrakcji danych AdvETL, które są koncepcyjnie i projektowo zasadniczo różne.

Klasy zaawansowanych hurtowni danych

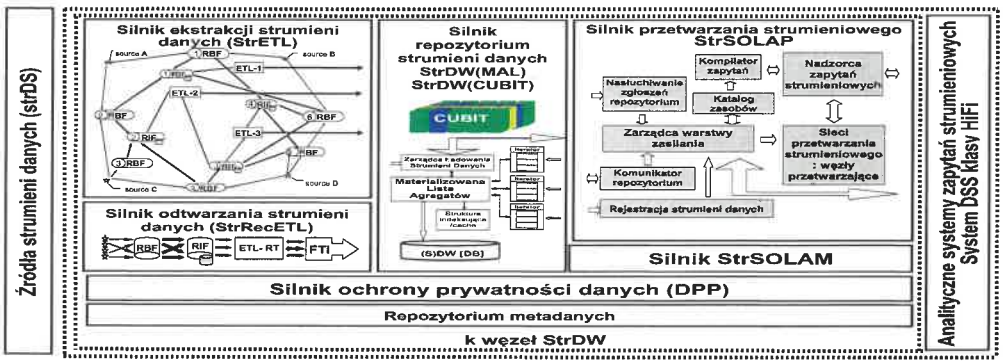
W III klasie AdvDW występują **Gridowe hurtownie danych (GDW^{SA})** z silnikami_ Agentami ekstrakcji danych strumieniowych (AgentAdvETL) oraz Agentami odtwarzania (AgentAdvRecETL) w każdym z n węzłów GDW^{SA}.

III klasa AdvDW: Gridowe hurtownie danych GDW^{SA}



W IV klasie AdvDW występują **Strumieniowe hurtownie danych (StrDW)** z konkretnymi specjalizowanymi silniki AdvETL takie jak: silnik ekstrakcji strumieni danych(silnik StrETL) oraz silnik odtwarzania strumieni danych (silnik StrRecETL).

IV klasa AdvDW: **Strumieniowe hurtownie danych StrDW**



Powyższy opis ramowy wskazuje zagadnienia badawcze w obrębie silników m.in. ETL i AdvETL z parowanych z silnikami odtwarzania procesów m.in. AdvRecETL i StrRecETL .

KONTEKSTY UTYLITARNO-EKSPERYMENTALNE SILNIKÓW ETL

Poziom technologiczny obecnie istniejących rozwiązań silników ETL częściowo zapewniają najważniejsze komercyjne narzędzia ETL i tak **klasyczne środowiska** to min.: Informatica, Talend, DataStage, Microsoft SQL Server Integration Services (SSIS), Oracle Data Integrator, Ab Initio oraz **zaawansowane środowiska** tj.: Airflow + dbt, Fivetran + dbt, Azure Data Factory, AWS Glue, Databricks). Generalne ETL dla systemów wielkiej skali to Informatica PowerCenter, IBM InfoSphere DataStage, SSIS, Apache NiFi, Apache Airflow, Talend oraz rozwiązania chmurowe tj AWS Glue czy Azure Data Factory. Istotne jest, aby wybrane środowisko ETL rozstrzygające

o efektywności całego procesu przetwarzania danych było dostosowane do indywidualnych potrzeb i możliwości danej organizacji/korporacji. Po tym wstępie można precyzyjnie określić tematykę rozprawy.

cd. Ocena celu i tezy rozprawy:
W dysertacji, na stronie 3 (1.3.1 Cele pracy) określono główne cele, a mianowicie:

1. Analiza formalna odległości edycyjnej drzew i grafów.
2. Opracowanie efektywnej metody wnioskowania pochodzenia danych.
3. Opracowanie narzędzia do mierzenia podobieństwa procesów ETL.
4. Implementacja i ewaluacja narzędzia do wnioskowania brakujących po wiązań pochodzenia danych.

Prezentowane cele mają uzasadniać dokonania badawczo-eksperymentalne w dość techniczny sposób przy braku sformułowanych tez.
Umiejscowienie dziedzinowe:
Rozprawa dotyczy BADAŃ PODSTAWOWYCH (jw.) w obrębie:

- Klasyczne systemy wspomagania decyzji podejmowania (systemy DSS) i klasyczne hurtownie danych (DW).
- Klasyczny silnik ETL (bez klasycznego silnika odtwarzania RecETL).

Wg Recenzenta temat rozprawy to zasadniczo rozłączne, nie uzupełniające się 2 cele i 2 tezy jak niżej.

Problem badawczy 1. ***Analiza podobieństwa procesów ETL .... w kontekście zarządzania hurtownią danych.***

Celem 1 rozprawy jest analiza podobieństwa procesów ETL z użyciem metody porównywania procesów ETL przez modelowane ich jako skierowane grafy acykliczne (DAG) i obliczanie odległości edycyjnej grafu GED (ang. Graph Edit Distance - GED) bazując na modelu procesów MetaDexa.
Doktorant nie sfomułował tezy w rozprawie. Można wyróżnić 2 tezy.
Teza 1. Możliwe jest ustalenie odpowiednich podobieństw procesów ETL z użyciem metody porównawczej do optymalizacji zadań ETL takich jak:

- identyfikacja redundancji procesów (duplikacja kodu),
- ułatwienia refaktoryzacji istniejących procesów,
- zwiększenia wsparcia ewaluacji procesów przez sugerowanie użycia gotowych fragmentów procesów.

Uzasadnienie celu 1 i tezy 1 to uzyskanie metod:

- formalnej analizy podobieństwa struktury procesów ETL na poziomie identyfikacji powtarzalnych wzorców,
- redukcji złożoności i eliminację nadmiarowości w procesach ETL,
- wsparcia automatycznej detekcji podobnych procesów ETL w dużych repozytoriach,

Problem badawczy 2. .. *pochodzenie danych w kontekście zarządzania hurtownią danych.*

Celem 2 rozprawy dla problemu badawczego 2 jest wg Doktoranta (rozdz. 1.3.1) wprost:

1. Opracowanie efektywnej metody wnioskowania pochodzenia danych na podstawie struktury schematów baz danych.
2. Implementacja i ewaluacja narzędzia do wnioskowania brakujących powiązań pochodzenia danych między tabelami i kolumnami.

Doktorant nie sfomułował także tezy 2 w rozprawie.

	Teza 2. Możliwe jest wnioskowanie o brakujących liniach danych z metadanych z użyciem modeli bazujących na transformatach w systemach analitycznych. ***Uzasadnienie celu 2 i tezy 2 to*** uzyskanie metody: • wnioskowania o brakujących powiązań pochodzenia danych, dzięki której możliwe staje się uzupełnianie i poprawa jakości danej informacji, • podległej walidacji na wielkich zbiorach danych, z którymi nie radzą sobie standardowe narzędzia skanujące oparte na analizie kodu lub logach. Problemem obu rozłącznych celów i tez jest zwiększenie jakości wiedzy o danych, a w konsekwencji wzrost zaufania do wyników analiz systemów DSS. Dokonanie oceny merytorycznej rozprawy umożliwi uzyskane odpowiedzi na dwa zasadnicze pytania: **1) Czy problem badawczy 1 oraz 2 ma znaczący charakter naukowy?** **2) Czy problem badawczy 1 i 2 ma duże znaczenie praktyczne?**
3	**Ocena merytoryczna rozprawy** Niniejszą rozprawę zaliczam do klasy prac badawczo-konstrukcyjno-wdrożeniowych. Nie ulega wątpliwości, że Doktorant osiągnął pewien stopień opanowania teorii inżynierii systemów i baz danych, a zwłaszcza wiedzy o badaniu, projektowaniu i prototypowaniu klasycznych systemów DSS i klasycznych systemów DW. Dysertacja jest skonstruowana dość poprawnie, zawiera zasadny wstęp oraz bardzo niejasne pojęciowo omówienie stanu wiedzy systemów DSS, systemów DW i silników ETL. Doktorant przedstawił główne cele rozprawy bez kontekstu badawczego przy braku sformułowania tezy. Z drugiej strony, utrudnieniem w czytaniu rozprawy była momentami duża nieprecyzyjność sformułowań technicznych i nie wprowadzeniu na początku pracy podstawowych pojęć i definicji. Stąd przedstawiono w recenzji powyższej (Rec. pkt. 2) kompendium wiedzy omawianej dziedziny dla uzyskania większej przejrzystości i czytelności tematyki rozprawy. **PROBLEM BADAWCZY 1 - JEGO CHARAKTER I ZNACZENIE** **Rozważmy realizację celu 1 i tezy 1 dotyczącej *problemu badawczego* 1. Analiza podobieństwa procesów ETL .... w kontekście zarządzania hurtownią danych.** W rozprawie pojawia się 2 - krotnie sformułowanie *zarządzania hurtownią danych* w tytule rozprawy i na stronie 17 w lakonicznym zadaniu: „„W rezultacie tradycyjne podejście do zarządzania hurtownią danych okazuje się niewystarczające". Doktorant wprowadzania w rozdz. 2.2.3 pojęcie zarządzanie danymi. Formalnie pod pojęciem zarządzania danymi rozumiemy modele zarządzania danymi, które definiują procesy, role i zasady zapewniające zgodność danych z wymaganiami prawnymi, operacyjnymi i jakościowymi. Skutecznie wdrażanie takich modeli w systemach DSS dla wielkich zbiorów danych wymaga głębokiej perspektywy różnych ich wymiarów min. tj. jakości danych, integracji danych z kilkuset źródeł danych, ekstrakcji danych równoległe wielkiej skali (tysiące węzłów transformacji), równoległe insertery pracujące w trybie superpozycji i indeksowania (np. UB-tree) wg specjalizowanych schematów danych w DW. Doktorant natomiast „skupia się na jego szczególnym wymiarze: formalnym nadzorze nad danymi i metadanymi, znanym jako *data governance*. To właśnie ten komponent zarządzania danymi odpowiada za definiowanie ról, procesów, zasad i odpowiedzialności w zakresie dostępu, jakości oraz wykorzystania danych w organizacji" (str.19). W rozdz. 2.4.2 doktorant określił pierwszy problem wyznaczenia podobieństwa procesów ETL-owych (także wielkoskalowe procesy ETL) w kontekście ich logiki

i przepływu danych, niezależnie od technologii, w której zostały zaimplementowane. Zdefiniowanie miary podobieństwa mają pozwalać zidentyfikować grupy procesów o zbliżonej logice i przepływie, dla wsparcia standaryzacji, konsolidacji kodu oraz decyzji biznesowych opartych na danych.

Problem podobieństwa procesów ETL-owych omówiono w rozdziale 5. Podstawy badań wcześniej zawiera rozdział 3 - analiza podobieństwa drzew. W rozdziale 4 Doktorant przedstawił metody porównywania grafów, traktowane jako uogólnienie analizy podobieństwa drzew, ze szczególnym uwzględnieniem odległości edycyjnej i z użyciem w analizie procesów ETL. Omówiono heurystyki przyspieszające obliczanie dokładnej wartości odległości edycyjnej oraz wersję przystosowaną do grafów DAG.

Kluczowe problemy zidentyfikowane w silnikach ETL zwłaszcza dla procesów ETL wielkich skali (zamiennie: wieloskalowe procesy ETL, ETL^{sl} , silniki ETL^{sl}) to a) rosnące koszty utrzymania ciągłości procesów, b) duże koszty wprowadzania zmian, c) wysoka złożoność obliczeniowa znalezienia dokładnego rozwiązania dla dużych grafów, d) różna granulacja realizacji operacji przetwarzania procesów.

Doktorant wykorzystał narzędzie MetaDex pełniącego rolę niezależnego skanera, który pozwala analizować dane w m.in. skomplikowane procedury ETL, dynamicznie generowany kod oraz transformacje realizowane w różnych językach i systemach. MetaDex, będący częścią rozwiązań platformy firmy Informatica, ładującego informacje o pochodzeniu danych do Cloud Data Governance and Catalog bazuje na patencie Pub. No.: US 2012/0296862 A1 autorstwa Dawid Duda, Zielonki (PL); Jeffrey T. Pascoe, Farmington, MI (US); Wojciech Matyjewicz, Tarnow (PL): Krzysztof Maziarz, Krakow (PL) pt METHOD AND APPARATUS FOR ANALYZING AND MIGRATING DATA INTEGRATION APPLICATIONS. Patent dotyczy metody migracji i analizy aplikacji do przenoszenia danych w oparciu o zebrane i przechowywane wszystkich związanych metadanych. Metadane opisują fizyczne i logiczne zasoby dane (tj. bazy danych, pliki danych, tabele bazy danych, kolumny bazy danych i ich typy danych), a także funkcje i parametry kosztowe przetwarzania. W opisie 0003 patentu podejście ETL wykorzystuje dedykowane platformy integracji danych, takie jak IBM Data Stage (dawniej Ascential), Informatica Power Center, Talend Open Studio/Talend Enterprise Data Integration itd. Aplikacja integrująca dane ma formę specjalnych programów (zwanych Jobs in Data Stage oraz Talend Open Studio/Talend Enterprise Data Integration, pingów map w Informatica itp.), zwykle realizowanych w formie graficznej, które opisują, skąd pobierać dane, jak je przetwarzać i gdzie je załadować.

MetaDex wprowadza grafowy model reprezentacji procesów ETL, umożliwiający odwzorowanie dowolnych procedur i procesów integracji danych niezależnie od użytej technologii. Dla każdej procedury ETL generowany jest oddzielny graf przedstawiający przepływ i transformacje danych. Łącząc grafy wszystkich procedur, otrzymujemy jeden całościowy graf obrazujący, jak dane przetwarzane są od źródła aż po finalne kolumny w systemie DW. MetaDex rozróżnia dwa typy zależności w grafie: a) wartościowe – wartość kolumny zależy od wartości kolumn poprzednich (np. w transformacjach, obliczeniach), b) kontrolne – istnieje zależność, przepływu danych, a nie ich wartości (np. w operacjach filtrowania, grupowania) (str.23). Doktorant dokonał szerokiej analizy formalnej tematyki podobieństwa drzew i grafów oraz silnych charakterystyk odległości edycyjnych GED (ang. Graph Edit Distance) (rozdz. 3 i 4). Ważność odległości edycyjnej w metodzie mierzenia podobieństwa struktur grafowych, wynika z definiowania jej jako minimalny koszt sekwencji operacji przekształcania jednego grafu w drugi. Doktorant zdefiniował matematyczny model MetaDex (def.5.1) oraz miara podobieństwa procesów ETL (def 5.2). To podobieństwo procesów ETL ma odzwierciedlać wpływ poszczególnych operacji na logikę transformacji i ich znaczenie dla danych wynikowych, przy zachowaniu struktury hierarchicznej w procesach ETL, czego nie zapewniały klasyczne algorytmy grafowe. Definicja 5.3 mówi o problemie identyfikacji podobnych procesów ETL. W rozdz. 5.6

doktorant przedstawił eksperymenty i wyniki. Dużym zaskoczeniem jest dobór 3 zbiorów danych, zanonimizowanych, obejmujących odpowiednio 1500, 7500 oraz 14000 procesów. Średnio, występuje mała liczba wierzchołków odpowiednio 41, 45 i 33, w pojedynczym procesie, i prawdopodobnie większość bardzo podobna, albo zupełnie inna. Doktorant upraszcza model MetaDex do prostego grafu z etykietami ograniczonymi do wyrażeń wartościowych i kontrolnych, ale tylko w obrębie operacji tego samego typu lub gdy jeden był podtypem drugiego. Następnie liczony jest GED za pomocą algorytmów bazowych A* + Hungarian oraz ich rozszerzenia autorskie z dwoma własnymi heurystykami Hungarian-A*-top (topologiczny porządek) i Hungarian-A*-cut (ograniczający wariant aproksymacyjny) Dodatkowe podejście autorskie DAG-Edit– podejście które wymusza zachowanie hierarchicznej struktury wynikającej z kierunkowości grafu. W eksperymentach losowano po 1000 grafów z każdego ww. zbioru danych. Grafy poddawano klasteryzacji, a następnie porównywano wszystkie pary w obrębie każdego klastra. Przykładowo, w tabeli 5.1 wykazano, że w testach dla zbioru A algorytm Hungarian-A-top* skrócił czas obliczeń z 22,5 s do 0,28 s w porównaniu do standardowego algorytmu A*, zachowując przy tym pełną optymalność wyniku przy koszcie 3.18. Doktorat wykazał w tym eksperymencie, że ww. heurystyki przyspieszają obliczenia i rzadko psują wynik. Powyższy eksperyment ma być dowodem realizowalności celu 1 i prawdziwości tezy 1 w formie weryfikacji koncepcji/metody jako typowy „proof-of-concept". Zatem idąc dalej należy zweryfikować cały eksperyment pod kątem zakresu stosowalności i użyteczności.
Teraz można przejść do udzielenia odpowiedzi na wcześniejsze pytanie:

Czy problem badawczy 1 ***(Analiza podobieństwa procesów ETL …. w kontekście zarządzania hurtownią danych)*** **ma znaczący charakter naukowy?**

W kontekście przedstawionego zarysu kompendium pt. BADANIA PODSTAWOWE i BADANIA ZAAWANSOWANE (pkt. 2 rec) **problem badawczy 1** *Analiza podobieństwa procesów ETL …. w kontekście zarządzania hurtownią danych)* **ma znaczący charakter naukowy.** Natomiast **problem badawczy 1** przedstawiony i opisany w dysertacji ma bardziej charakter badawczo- techniczny, a **charakter naukowy znacznie pomniejszony** przez pewne wyniki teoretyczne wybiegające poza obszar naukowy *zarządzania hurtownią danych.* Ciekawym wynikiem jest formalne wykazanie braku podliniowej redukcji instancyjnej między odległością edycyjną a wyrównaniem drzew (rozdz. 3.5, tw.3.45). Doktorant udowodnił, że problem wyznaczania największego wspólnego poddrzewa oraz problem wyznaczania najmniejszego wspólnego naddrzewa, prowadzą do zasadniczo różnych rezultatów i koryguje wcześniejsze ustalenia zawarte w literaturze (formalny kontrprzykład do relacji porządku).

OCENA 1.1. (znaczący charakter naukowy)

Problem badawczy 1 ***(Analiza podobieństwa procesów ETL …. w kontekście zarządzania hurtownią danych)*** **ma znaczący charakter naukowy** w dyscyplinie naukowej Inżynieria Danych (Bazy Danych/Hurtownie Danych). Rozprawa dotyczy BADAŃ PODSTAWOWYCH (jw.) w obrębie: a) Klasyczne systemy wspomagania podejmowania decyzji (DSS) i klasyczne hurtownie danych (DW), b Klasyczny silnik ETL bez klasycznego silnika odtwarzania RecETL. Natomiast **problem badawczy 1** przedstawiony i opisany w dysertacji ma bardziej charakter badawczo- techniczny, a **charakter naukowy znacznie pomniejszony** mimo uzyskanych pewnych wyników teoretycznych wybiegających poza obszar naukowy *zarządzania hurtownią danych.*
Dalej, wadą opisu jest jego niekiedy enigmatyczny język z wieloma niewyjaśnionymi terminami, co na pewno utrudni lekturę tekstu osobie niebędącej ekspertem w dziedzinie. Teraz można przejść do udzielenia odpowiedzi na pytanie drugie:

Czy problem badawczy 1 *(Analiza podobieństwa procesów ETL .... w kontekście zarządzania hurtownią danych)* **ma duże znaczenie praktyczne?**
Do oceny **znaczenia praktycznego** wyróżniamy dwie **grupy użyteczności modelu MetaDex z hurystykami** tj.: **Grupy stosowalności (GS)** oraz **Grupy przydatności praktycznej (GU).**
Grupy stosowalności (GS) ww. eksperyment traktowane jako ograniczenia.
GS 1. Ograniczenia grafów ETL min.:

- do jednego insertera ładującego,
- o rozmiarze do 20 wierzchołków (losowe pary grafów bez wcześniejszej klasteryzacji ze względu na wysoką złożoność czasową),
- o rozmiarze do 50 wierzchołków przy metodzie klasteryzacji.

GS 2. Ograniczenia projektowo-testowe min. tj.:

- max 1000 losowych par oraz 1000 par z tego samego klastra na 14 000 procesów,
- uproszczone i homogeniczne procesy ETL,
- mała skala i granularność grafów/procesów ETL,
- funkcje kosztów nie walidowane i nieuzasadnione tj. (np. koszt usuwania krawędzi =10, koszt dopasowania wierzchołków =2 dla algorytmu Hungarian-A*-cut)

Grupy użyteczności (GU) ww. eksperymentu traktowane jako przydatność praktyczna.
GU 1. Przydatność praktyczna dla grafów ETL
Doktorant nie pokazał ani jednego przykładu podobieństwo dla grafów ETL min.:

- z setkami wierzchołków i krawędzi (> 500 wierzchołków) w realistycznej warstwie transformacji danych (WTD) klasycznych silników ETL,
- z tysiącami wierzchołków w WTD zaawansowanych silników StrETL,
- z wielokrotnymi rozgałęzieniami i równoległymi ścieżkami,
- z wieloinsertorowymi węzłami ładującymi,
- z wieloma podgrafami w warstwach realnych subprocesów AdvETL,
- z węzłami odtwarzania ETL w silnikach ETL oraz AdvETL.

GU 2. Przydatność praktyczna dla projektowania i testowania grafów ETL .
Doktorant nie pokazał praktycznego eksperymentu walidacji użytkowo-biznesowej dla grafów ETL, w szczególności realizacji tezy 1. Nie pokazano eksperymentu na potwierdzenie o uzyskaniu przydatnej metody pomocnej przy min.:

- refaktoryzacji kodu i wykrywaniu duplikatów,
- redukcji złożoności i eliminacji nadmiarowości w procesach ETL,
- porównaniu z ekspercką oceną podobieństwa grafów ETL,
- detekcji automatycznej podobnych procesów ETL w dużych repozytoriach,
- zrealizowanego w innym środowisku niż Informatica Cloud Data Integration i MetaDex.

OCENA 1.2. (duże znaczenia praktyczne)
W kontekście Grupy stosowalności (GS1 i GS2) eksperymentu przedstawionego w rozdziale 5.6 traktowane jako ograniczenia wskazują że waga uzyskanych wyników ma **bardzo umiarkowane znaczenia praktyczne.**
Warunkiem wystarczającym uznania, że problem badaczy 1 w może mieć w przyszłości **duże znaczenie praktyczne** jest realizacja Grupy użyteczności (GU1 i GU2) traktowane jako przydatności praktyczne.

Konkluzje i wskazania do celu 1 i tezy 1.
W obliczu powyższej skróconej analizy warunków: Ograniczenia GS1 i GS2 oraz Przydatność praktyczna GU1 i GU2 nazbyt ogólne są wnioski wysnute w podsumowaniu (rozdz. 5.7). A mianowicie Doktorant znacznie pomniejsza znaczenie tego, że w pracy nad omawianą metodą napotkano trudności:

- związane z dużą heterogenicznością procesów ETL (różnorodność systemów, konwencji nazewniczych i typów transformacji sprawiała, że bezpośrednie porównanie schematów jest problematyczne),
- w zachowaniu skalowalności metody przez zastosowanie mechanizmów unifikacji grafów oraz wstępnego klasteryzacji procesów.

Na wyrost także jest podobne stwierdzenie Doktoranta, że przedstawiona metoda i eksperyment pozwalają:

- wspierać optymalizację i standaryzację przepływów danych używając uniwersalnej miary podobieństwa procesów i może być wykorzystana do identyfikacji powtarzalnych wzorców i duplikatów,
- wspomagać projektowanie nowych procesów ETL, podpowiadając efektywne wzorce i minimalizując ryzyko powielania błędów lub nieefektywnych rozwiązań.

Generalnie, obecne wady wdrożenia praktycznego modelu MetaDex i heurystyk zgodne z tematykę celu 1 i tezy 1 to:

- Wysokie zagrożenie braku skalowalności dla dużych grafów ETL przy dziesiątkach tysięcy procesów z setkami węzłów i krawędzi przy niewydolności obliczeniowej omawianych heurystyk.
- Nieakceptowalny, duży koszt standardowych obliczeń GED jako problemu NP-trudnego, a mało znane są strategie szybkich obliczeń GED dla zbiorów wielkich rozmiarów.
- Nieoptymalne heurystyki powinny być zastąpione przez nowoczesne modele uczenia podobieństwa grafów dużej skali w kontekście metod analizy danych i sztucznej inteligencji.
- Duże uproszczenie semantyczne modelu MetaDex, gubiąc koncepcję uniwersalności zapisów patentu Pub. No.: US 2012/0296862 A1.
- Nie uwzględnienie analizy podatności na zmianę parametrów i walidacji kosztów rzeczywistej refaktoryzacji i utrzymania kodu.
- Nieakceptowalny brak walidacji z ekspertami ETL ani porównania z innymi hybrydowymi metrykami BPM.
- Brak obliczenia wskaźnika ROI uwzględniającego koszty wdrożeniowe przy redukcji różnej złożoności algorytmicznej i określonej precyzji modelu MetaDex/Heurystyki.
- Ograniczenie wdrożenia w rzeczywistym środowisku przemysłowym wyłącznie do użycia stosu technologicznego Informatica.

Koncepcję i wyniki opublikowano w artykule autorstwa:

1. Maciej Brzeski i Adam Roman. "Measuring Similarity Between ETL Processes Using Graph Edit Distance". W: Schedae Informaticae 32 (2023).
2. Maciej Brzeski. „A counterexample to linear relationship between largest common subtrees and smallest common supertrees". 2025. arXiv: 2509.12360 [math.CO]. URL: https://arxiv.org/abs/2509.12360.
3. Maciej Brzeski. A Counterexample to the Order Relation Between Largest Common Subtree and Smallest Common Supertree. W: Annals of Combinatorics. (2026). https://doi.org/10.1007/s00026-025-00803-9.

PROBLEM BADAWCZY 2 - JEGO CHARAKTER I ZNACZENIE

Rozważmy realizację celu 2 i tezy 2 dotyczącej problemu badawczego ***2. ..... pochodzenie danych w kontekście zarządzania hurtownią danych.***

Klasyczne metody rekonstrukcji pochodzenia danych to statyczna analiza kodu i ewaluacja w czasie działania ograniczone, gdy linie pochodzenia danych (ścieżki i transformacje) obejmują zróżnicowane systemy źródłowe oraz ich niedostępność prawna (min. ochrona prywatności danych). Wobec tych ograniczeń klasycznych

metod rekonstrukcji pochodzenia danych Doktorant proponuje podejście alternatywne - wnioskowanie o pochodzeniu danych na podstawie dostępnych metadanych schematu. Podstawowym obiektem analizy są pary schematów, obejmujące schemat źródłowy i schemat docelowy (z których każda często zawiera po kilka tysięcy kolumn). W rozdziale 7.4 mamy ogólny dwustopniowy model oparty o architekturę Transformera i obejmuje dopasowywanie kolumn poprzez segmentację, a następnie etapy filtrowania i klasyfikacji. Ten model typu Transformer łączy wydajność bi-kodera z precyzją kodera krzyżowego. Tu etap filtrowania oparto na bi-koderze, który niezależnie koduje kolumny źródłowe i docelowe, generując osadzenia wektorowe dla obliczeń na GPU. Bi-koder trenowano z użyciem funkcji strat MNR. Drugi etap wykorzystuję koder krzyżowy, który wspólnie koduje pary kolumn i klasyfikuje je jako powiązane bądź nie. W testach praktycznych użyto warianty modelu MiniLM (rodzina ms-marco-MiniLM-L#-v2), dla osiągniecia akceptowalnego kompromisu między rozmiarem a skutecznością.

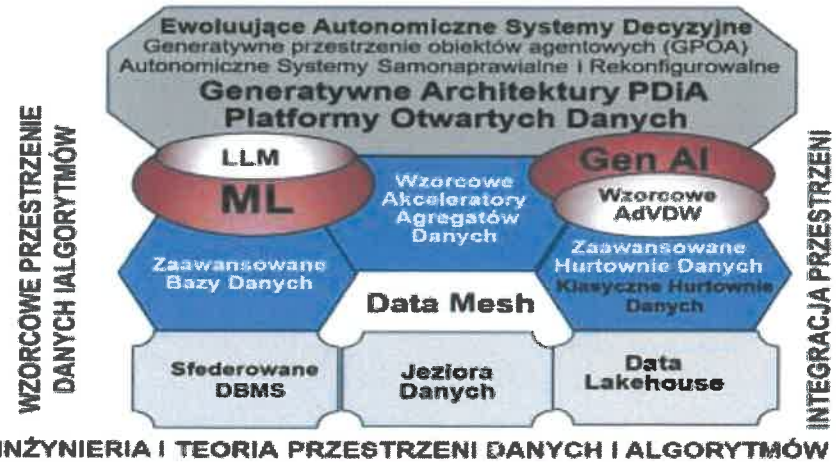
Dokonanie oceny merytorycznej rozprawy umożliwi uzyskane odpowiedzi na dwa zasadnicze pytania:

2.1) Czy problem badawczy 2 (.... pochodzenie danych w kontekście zarządzania hurtownią danych) **ma znaczący charakter naukowy**?

2.2) Czy problem badawczy 2 (.... *pochodzenie danych w kontekście zarządzania hurtownią danych)* **ma duże znaczenie praktyczne?**

OCENA 2.1. (znaczący charakter naukowy)

Śledzenie pochodzenia danych odgrywa kluczową rolę w zarządzaniu hurtownią danych. Wiarygodne dokumentowanie przepływu danych, od źródeł, przez różne transformacje warstwowe, aż do końcowych wyników zapewnia organizacjom ważne informacje o pochodzeniu, jakości, przejrzystości, zależnościach i konsekwencjach użycia tych danych. W praktyce jednak śledzenie pochodzenia danych jest zwykle mocno utrudnione lub ograniczone przez różne czynniki. Często w dużych organizacjach (przemysł, banki, ubezpieczenia) brakuje instrumentacji, systemów dziedziczonych, developmentu Pythona/Java bez parsowalnego kodu. Generalnie, należy metodycznie wspomagać się zaawansowaną strategią zarządzania danymi w systemach transakcyjnych i systemach analitycznych. Taką ukierunkowaną na zarządzanie hurtownią danych jest **strategia zarządzania Przestrzenią Danych i Algorytmów (PDiA)**.



Analityka danych bazuje na osiągnięciach inżynierii danych, wykorzystując dane do predykcji, decyzji i obserwowalości. Klasyczną analitykę danych tworzą dwie warstwy tj.: a) analizy, metryki, segmenty, zagregowane danych i funkcje zorientowane oraz b) sztuczna inteligencja i głębokie uczenie. W wykładzie pod linkiem Wykład Profesorski - prof. dr hab. inż Marcin Gorawski przedstawiono nową strategię zarządzania PDiA.

Tak ulokowany **problem badawczy 2** (.... *pochodzenie danych w kontekście zarządzania hurtownią danych)* **ma znaczący charakter naukowy**.

Doktorant przedstawił metodę zwinną wykorzystującą dwustopniowe podejście NLP o architekturze transformera do wnioskowania o pochodzeniu danych wyłącznie z użyciem metadanych. Metoda nie wymaga dostępu do danych ani kodu – tylko nazwy kolumn. Doktorant wskazuje ograniczenia tej metody wnioskowania pochodzenia danych i zalicza do nich:

- Uwzględnienie wyłącznie dostępnych metadanych, a więc tylko nazw tabeli i nazw kolumn _po segmentacji, żadnych danych instancyjnych typów danych, statystyk kolumn, łączeń warunkowych, wyrażeń obliczeniowych, kontekstu grafu.
- Uproszczenia tylko do transformacji jednoźródłowych (obserwowanych takich mapowań rzeczywistym zbiorze danych to 95,3%, pozostałe 4,7% to mapowania wieloźródłowe).
- Brak indywidualnej metody podejścia przy każdej implementacji, co generuje trudności ze skalowalnością i niejednorodnością metadanych w środowiskach produkcyjnych.

OCENA 2.2. (duże znaczenia praktyczne)
Tu podobnie jak w ocenie 1.2 eksperyment (rozdz. 7.3 i 7.5) ma być dowodem realizowalności celu 2 i prawdziwości tezy 2 w formie weryfikacji koncepcji/metody jako typowy „proof-of-concept". Zatem idąc dalej należy zweryfikować cały eksperyment ze względu na w**arunki eksperymentu**

- **Zbiór danych**:
- Skalę zbioru danych opisuje ponad 47 milionów kolumn z 77 237 schematów i 5 673 różnych źródeł.
- Zakres analizy obejmuje poddano 57 546 par schematów i około 16 milionów zweryfikowanych powiązań pochodzenia danych (na każde około 90000 możliwych par przypada tylko jedno pozytywne powiązanie).
- Zrekonstruowaną strukturę baz danych (tabele, schematy) na podstawie niepełnych logów transformacji.
- **Warunki eksperymentu**
- Metoda podziału danych na zbiór treningowy (90%) i testowy (10%) przy podejściu międzyprojektowym (bezwzględna separacja między schematami).
- Treningi przeprowadzono na komputerze z kartą NVIDIA A100 z użyciem dostrojonych modeli opartych na architekturze BERT.
- Czas trwania pojedyncza sesji treningowej to około 10 godzin dla bi-kodera oraz 7 godzin dla kodera krzyżowego, 40h generowanie trudnych negatywów.
- Użyto miarę skuteczności klasyfikacji PR-AUC ze względu na silnie niezrównoważony charakter zbioru danych.

Uzyskane wyniki eksperymentu
Użyto miarę skuteczności klasyfikacji PR-AUC (ang. Precision-Recall Area under Curve) ze względu na silnie niezrównoważony charakter zbioru danych.

- Zbadano, jak próg podobieństwa bi-kodera przy wyborze trudnych negatywów (HN) wpływa na skuteczność modelu kodera krzyżowego. Tabela 7.1. Najlepszą skuteczność koder krzyżowy osiągnął dla progu HN: 0,4 (optymalny próg trudnych negatywów), przy najwyższej wartości miary AUC (0,82) oraz precyzji (0,74) i czułości (0.81).
- Zbadano wpływ liczby przykładów negatywnych na skuteczność modelu. Zwiększanie liczby trudnych przykładów powyżej n=2 ma ograniczony wpływ na skuteczność. Tabela 7.2 pokazuje, że miara F1 stabilizuje się na poziomie 0,78 dla wszystkich wartości n od 2 do 5.

- Z analizy ablacji wynika, że usunięcie segmentacji lub trudnych negatywów drastycznie obniża jakość modelu. Tabela 7.3 pokazuje spadek AUC z 0,80 (pełny model) do odpowiednio 0,65 (model bez segmentacji) oraz 0,61 (model bez trudnych negatywów).
- Badania procesu trenowania modelu na danych z wielu organizacji/firm daje lepsze wyniki niż ograniczenie się do jednej organizacji/firmy, co potwierdzają wyniki tabeli 7.4 to dla każdego zbioru (zanonimizowanych A, B, C) – np. dla zbioru B AUC w podejściu międzyprojektowym wynosi 0,99 (Precyzja 0.99, Czułość 0.95) wobec 0,79 w wewnątrzprojektowym (Precyzja 0.76, Czułość 0.92).
- Określono wpływ wielkości modelu na skuteczność kodera krzyżowego, testując różne warianty modelu MiniLM i w tabeli 7.5 widzimy, że poza najmniejszym wariantem (L2-v2), skuteczność modeli jest zbliżona (AUC na poziomie 0,79-0,80 dla modeli od L4 do L12, podczas gdy L2 osiąga jedynie 0,74.

Konkluzje i wskazania do celu 2 i tezy 2.
Doktorant przedstawił metodę wnioskowania (brakujących powiązań) pochodzenia danych przy użyciu dwustopniowego modelu o architekturze transformera. Eksperymenty potwierdziły skuteczność i celowość zastosowania takiej dwustopniowej architektury w testach na rzeczywistych danych. Ważna jest walidacja modelu uwarunkowana:

- Zasadą bezwzględnej separacji między schematami, celem uniknięcia „przecieku informacji" (wszystkie dane z konkretnej firmy trafiały albo do zbioru treningowego, albo testowego, nigdy do obu).
- Procedurą ablacji, która pokazuje, że testy usuwając poszczególne komponenty (segmentację, trudne negatywy), pokazują, że każdy z nich realnie wpływa na poprawę jakości, a nie jest tylko zbędnym dodatkiem.
- Mechanizmem generowania trudnych negatywów i procedura segmentacji nazw technicznych, co pozwala modelom LLM operować na danych, do których nie były pierwotnie trenowane.
- Użyciem silnej zasady walidacji międzyprojektowej gwarantującej rzeczywistą zdolność modelu do generalizacji.

Mocna strona rozwiązania problemu badawczego 2

- Utylitarna motywacja ze wskazaniem praktycznego rozwiązania.
- Model dwustopniowy (I stopień to szybki filtr kandydujący bi-koder)
- Dostępny kod źródłowy oraz syntetyczny zbiór danych.
- Silna walidacja modelu (generalizacja międzyprojektowa).
- Rzetelnie opisane ograniczenia modelu i metody wnioskowania.
- Poprawna analiza ablacji dla segmentacji i trudnych negatywów.
- Podejście komplementarne do rozwiązania problemu badawczego 1.
- Koncepcję i wyniki opublikowano w artykule autorstwa: Maciej Brzeski i Adam Roman. "*Inferring Missing Data Lineage Links from Schema Metadata Using Transformer-Based Models*". W: Proceedings of the VLDB Workshop 2025: 6th Applied AI for Database Systems and Applications ISSN (2025).

Słaba strona rozwiązania problemu badawczego 2 w kontekście braku:

- odniesienia do rozwiązania problemu badawczego 1,
- porównania z metod grafowymi,
- benchmarków dla schematów z setkami tysięcy kolumn (np. 700 000),
- procesu skalowalności ogromnych schematów w jednym systemie,
- zaawansowanej architektury modelu uwzględniającej min.: duże modele LLM, wstępnego trenowania danych domenowych, użycie DAPT (ang. Domain-Adaptive Pre-Training),

	• nowoczesnych architektur enkoderowych (np. DeBERTa-v3, E5-Mistral 7b), ani technik late-interaction (np. ColBERT-v2), • walidacji krzyżowej ze względu na wysoki koszt obliczeniowy (sam trening bi-kodera to 10h, a generowanie negatywów 40h).
4	**Wkład naukowo-wdrożeniowy Doktoranta opisywany w rozprawie** Wkład naukowo – wdrożeniowy doktoranta mgr Macieja Brzeskiego należy postrzegać jako znaczący udział we współpracy dwóch Instytucji tj. Katedry Inżynierii Oprogramowania w Instytucie Informatyki i Matematyki Komputerowej Wydziału Matematyki i Informatyki UJ i firmy Informatica. Stopień zaangażowania i szczegółowy wkład naukowo – wdrożeniowy doktoranta najlepiej może określić jego promotor prof. dr hab. Adam Roman, kierownik Katedry Inżynierii Oprogramowania oraz opiekun pomocniczy mgr Dawid Duda. Dysertacja ma charakter doktoratu wdrożeniowego, stąd mierzalność efektów była weryfikowana w rzeczywistym środowisku zintegrowanym z komercyjnymi produktami firmy Informatica (MetaDexa, Informatica Cloud Data Governance and Catalog (CDGC)). Temat rozprawy doktorskiej to zasadniczo rozłączne dwa cele i dwie tezy (Rec.str.6) **Wkład naukowo-wdrożeniowy w problemie badawczym 1**. *Analiza podobieństwa procesów ETL …. w kontekście zarządzania hurtownią danych.* Doktorant potwierdza, że narzędzia zostały pomyślnie zaimplementowane i ewaluowane w skali przemysłowej, co pozwoliło na realne usprawnienie monitorowania procesów ETL oraz wzbogacenie informacji o pochodzeniu danych. Wysoką jakość wdrożenia zapewnia fizyczny poziom pochodzenia danych, tu przyjęto poziom tabel i kolumn jako dokładne odwzorowanie przepływu danych między obiektami i wszystkimi transformacjami tj. agregacje, filtrowania, łączenia itp.). MetaDex jest częścią rozwiązań firmy Informatica, pełni rolę skanera zintegrowanego z platformą Informatica, ładującego informacje o pochodzeniu danych do CDGC, zapewniając spójny i całościowy widok pochodzenia danych dla całego przedsiębiorstwa. Poza tym platforma Informatica daje udokumentowane oraz certyfikowane pochodzenie, akceptowalne przez audytorów. Bazując na modelu MetaDex/GED doktorant przedstawił eksperymenty i wyniki (rozdz. 5.6). Dobrano 3 zbiory danych, zanonimizowanych, obejmujących odpowiednio 1500, 7500 oraz 14000 procesów. Średnio, występuje mała liczba wierzchołków odpowiednio 41, 45 i 33, w pojedynczym procesie. I tutaj doktorant upraszcza model MetaDex do prostego grafu z etykietami ograniczonymi do wyrażeń wartościowych i kontrolnych, ale tylko w obrębie operacji tego samego typu lub gdy jeden był podtypem drugiego. Następnie liczony jest GED za pomocą algorytmów bazowych A* + Hungarian oraz ich rozszerzenia autorskie z dwoma własnymi heurystykami Hungarian-A*-top (topologiczny porządek) i Hungarian-A*-cut (ograniczający wariant aproksymacyjny). Przykładowy wymierny zysk to policzalny wzrost wydajności dla zbioru A algorytmu H-A* (22.5s) względem algorytmu H-A*-top (0,28s) (np. skrócenie czasu obliczeń z 22s do 0,28s) potwierdza inżynierską kompertencji (Tab. 5.1). zachowując przy tym pełną optymalność wyniku przy koszcie 3.18. W recenzji na stronie 10 (Rec.str.10) oceniono **znaczenie praktyczne** wyróżniamy dwie **grupy użyteczności modelu MetaDexa z Hurystykami** tj.: **Grupy stosowalności (GS)** oraz **Grupy przydatności praktycznej (GU)** (Rec.str. 10). Z Oceny 1.2 (w kontekście Grupy stosowalności (GS1 i GS2) eksperymentu traktowane jako ograniczenia wskazują że waga uzyskanych wyników ma **bardzo umiarkowane znaczenia praktyczne.** Generalnie, obecne wady wdrożenia praktycznego modelu MetaDex/heurystyki zgodne z tematykę celu 1 i tezy

	wymieniono szczegółowo w **Konkluzjach i wskazaniach do celu 1 i tezy 1** (Rec.str. 11). **Wkład naukowo-wdrożeniowy w problemie badawczym 2.** ... *pochodzenie danych w kontekście zarządzania hurtownią danych.* Doktorant przedstawił metodę wnioskowania (brakujących powiązań) pochodzenia danych przy użyciu dwustopniowego modelu o architekturze transformera (I stopień to szybki filtr kandydujący bi-koder; II stopień to koder krzyżowy zapewniający maksymalną precyzję predykcji linku). W niniejszej recenzji (Rec,str.13) przedstawiono szczegółową **ocenę 2.2.** (duże znaczenia praktyczne) oraz analizę **uzyskanych wyników eksperymentu**. Z punktu widzenia samego wdrożenia tego modelu Doktorant wykazał, że istotne jest: • zastosowanie trudnych przykładów, wyselekcjonowanych przez bi-koder, jest kluczowe dla skuteczności w rzeczywistych systemach, • uczenie na "trudnych negatywach" - modele uczone na „łatwych negatywach" (losowych danych) zawodzą w rozróżnianiu kolumn o podobnych nazwach, a różnym znaczeniu, • użycie segmentacji tekstu o nazwie *Stupid Backoff* tj. prostej i skalowalnej metody częstości występowania zamiast złożonych modeli ML dla podziału nazw kolumn jest uzasadniony specyfiką tekstu (skróty, brak separatorów), gdzie kontekst n-gramowy jest wystarczający, • odejście od danych instancyjnych, co uzasadnia oparcie modelu wyłącznie na metadanych (nazwach schematów/kolumn) w realiach utylitarnych, gdzie dostęp do właściwych danych jest często niemożliwy ze względów na bezpieczeństwa i regulacje prawne, • wyjście poza nieliczne i syntetyczne zbiory często spotykane w pracach akademickich, do wielkich zbiorów dane pochodzących z autentycznych systemów dużych instytucji (banki, ubezpieczalnie), • testowanie na całkowicie obcych konwencjach nazewniczych wg zasady „separacji projektów/firm" jako silna strona walidacji - dla uniknięcia „przecieku informacji", wszystkie dane z konkretnej firmy trafiały albo do zbioru treningowego, albo testowego, nigdy do obu, • zastosowanie bi-kodera w celu skutecznego filtrowanie na niemal 100-krotne mniejszej liczbie kandydatów do dopasowania w porównaniu do np. współdzielenia tokenów, • wykorzystania maszyny z dwiema kartami graficznymi NVIDIA A100 do trenowania i wnioskowania na dużą skalę (np. zbiór 47 milionów kolumn), • użycia środowiska dewelopmentu języka Python, wykorzystując biblioteki do uczenia głębokiego - kod źródłowy oraz dane syntetyczne zostały udostępnione w serwisie GitHub. W rozdziale **Konkluzje i wskazania do celu 2 i tezy 2.** (Rec.str. 14) przedstawiono • *Mocną stronę rozwiązania problemu badawczego 2,* • *Słabą stronę rozwiązania problemu badawczego 2 .* Doktorant nie ukrywa, że jego metoda wnioskowania jest formą "uzupełniania luk", gdzie tradycyjne skanery zawodzą, ale także przyznaje, że jego metoda nie sprawdza się w każdym przypadku i scenariuszu.
5	**Nowa Architektura Modelu MetaDex/GDE/2EnKodery** Doktorant w końcowym rozdziale. 8.1 stwierdził: *„Celem niniejszej rozprawy było zbadanie dwóch powiązanych zagadnień w kontekście hurtowni danych: analizy podobieństwa procesów ETL oraz wnioskowania brakujących elementów pochodzenia danych".*

W niniejszej recenzji wykazano wprost, że takie powiązanie nie zachodzi. W dysertacji przedstawiono rozwiązania dwóch rozłącznych problemów badawczych. Temat rozprawy to zasadniczo rozłączne dwa cele i dwie tezy jak niżej.
Problem badawczy 1. ***Analiza podobieństwa procesów ETL .... w kontekście zarządzania hurtownią danych.***
Problem badawczy 2. .. *pochodzenie danych w kontekście zarządzania hurtownią danych.*
W recenzji wykazano, że pojęcie „***zarządzania hurtownią danych***" jest tu niedookreślone i stąd nazbyt luźne rozumienie struktury grafów ETL i procesów ekstrakcji danych realizowanych w silnikach ETL, nawet tych w klasycznych systemach hurtowni danych.

W **rozwiązaniu problemu badawczego 1** użyty został model MetaDex bazujący na koncepcji patentu Pub. No.: US 2012/0296862. Pełny model MetaDex opisuje semantykę metadanych (w tym struktury obliczeń i zależności między kolumnami) zapewniając wartościową reprezentację schematów źródeł. MetaDex to model strukturalny - może wykorzystywać rozkłady wartości, typy danych i zależności statystyczne. MetaDex to narzędzie pełniące rolę niezależnego skanera, nie przywiązanego do konkretnego katalogu danych. Dla każdej procedury ETL generowany jest oddzielny graf przepływu i transformacji danych. W dysertacji natomiast zastosowano uproszczenie semantyczne modelu MetaDex, do prostego grafu z etykietami ograniczonymi do wyrażeń wartościowych i kontrolnych, ale tylko w obrębie operacji tego samego typu lub gdy jeden był podtypem drugiego. Następnie liczony jest GED za pomocą algorytmów bazowych A* + Hungarian oraz z dwoma własnymi heurystykami Hungarian-A*-top (topologiczny porządek) i Hungarian-A*-cut (ograniczający wariant aproksymacyjny). Stąd mamy środowisko rozwojowe MetaDex/GED lub zamiennie MetaDex/heurystyki.

W **rozwiązaniu problemu badawczego 2** zbudowano dwustopniowy model wnioskowania o pochodzeniu danych. W praktyce ten model sprowadzony jest zagadnień tokenizacji, segmentacji i podobieństwa nazw. W konsekwencji do porównania nazw kolumn po segmentacji.

Poniżej Recenzent prezentuję pewien szkic architektury spójnej dla tematu rozprawy.
Nowa Architektura Modelu MetaDex/GDE/2EnKodery.
Generalnie w celu połączenia tych 2 rozwiązań proponuję rozważyć przykładową architekturę z użyciem:

1. Odległości edycyjnej grafu (GED) bazując na modelu procesów MetaDexa (MetaDex/GED) jako pewnych i zaufanych etykiet/krawędzi do trenowania lub oceniania modelu wnioskowania 2EnKodery. Wtedy można przyjąć dla grafu a) brak krawędzi to negatyw, b) występowanie krawędzi to pozytyw, c) typy krawędzi to rodzaj relacji, c) ścieżki w grafie to semantyka.
2. Bootstrapowania grafu ETL dla niepełnego graf z brakującymi krawędziami z wykorzystaniem modelu wnioskowania 2EnKodery. Ogólnie, bootstrapowanie grafu ETL to proces, w którym model ML uzupełnia brakujące krawędzie w grafie, a graf ETL dostarcza modelowi nowych etykiet w cyklu iteracyjnym.

Istnieje jednak taki najgorszy możliwy przypadek gdy nie znamy: grafu ETL, kodu źródłowego ETL, wyrażeń, zależności, semantyki i struktury a istnieje tylko informacja o nazwach kolumn w schematach źródeł danych. Wtedy dwupstopniowy model wnioskowania może być wdrażalny w każdej firmie, natomiast model MetaDex/GED wymaga wdrożenia kosztownej infrastruktury operacyjnej.

6	**Dane o dorobku naukowym Doktoranta**

Poniżej tabelarycznie przedstawiono dane o dorobku naukowym Doktoranta.

Rodzaj publikacji	Liczba
Artykuły w czasopiśmie	4
Monografie naukowe	0
Rozdziały w monografiach	0
Publikacje konferencyjne	3
Pozostałe publikacje	2
Razem publikacji	9
Punktacja MNiSW	210

Publikacje wyróżnione	Liczba	Pozycja poniżej
Artykuły w czasopiśmie z IF	1	[1] (IF 0.7 (2024)(100)
Artykuły w czasopiśmie UJ	3	[4](20),[6](11),[8](11)
Publikacje_konferencyjne (liczone 140 pkt. MNiSW)	0	
Publikacje_konferencyjne (pierwszy współautor)	2	[5](70),[3](0)
Publikacje_konferencyjne (Inne)	1 2	[7](0) [2](0),[9](0)

1. Maciej Brzeski. ***A Counterexample to the Order Relation Between Largest Common Subtree and Smallest Common Supertree***. W: Annals of Combinatorics. (2026). https://doi.org/10.1007/s00026-025-00803-9 (IF 0.7 (2024)), (pkt. MNiSW 100).
2. Maciej Brzeski. ***A counterexample to linear relationship between largest common subtrees and smallest common supertrees***. 2025. arXiv: 2509.12360 [math.CO]. URL: https://arxiv.org/abs/2509.12360. (pkt MNiSW. 0).
3. Maciej Brzeski i Adam Roman. ***Inferring Missing Data Lineage Links from Schema Metadata Using Transformer-Based Models***. VLDB Workshops – Applied AI for Database Systems and Applications. W: Proceed of the VLDB Endowment. (2025), (pkt MNiSW. 0).
4. Maciej Brzeski i Adam Roman. ***Measuring Similarity Between ETL Processes Using Graph Edit Distance***. Wydawnictwo UJ, Schedae Informaticae Journal, vol.32 (2023), (pkt. MNiSW 20).
5. Maciej Brzeski, Hubert Dryja, Paweł Gora, Karnas Katarzyna, Arkadiusz Klemenko, Adrian Kochański, Dawid Kopczyk, Magdalena Kukawska, Marcin Możejko, Przemysław Przybyszewski. ***Solving Traffic Signal Setting Problem Using Machine Learning.*** International Conference On Models and Technologies for Intelligent Transportation Systems (2019), (pkt. MNiSW 70).
6. Możejko Marcin, Brzeski Maciej , Mądry Łukasz , Skowronek Łukasz , Gora Paweł. ***Traffic Signal Settings Optimization Using Gradient Descent***. Theoretical Conference of Machine Learning, Wydawnictwo UJ, Schedae Informaticae Journal vol.27. DOI 10.4467/20838476SI.18.002.10407 (2018), (pkt. MNiSW N 11).
7. Pawel Góra, Maciej Brzeski, Marcin Mozejko, Arkadiusz Klemenko, Adrian Kochanski. ***Investigating performance of neural networks and gradient boosting models approximating microscopic traffic simulations in traffic optimization tasks.*** NeurIPS Workshop – Machine Learning for Intelligent Transportation Systems, 32nd Conference on Neural Information Processing Systems (NIPS 2018), (pkt MNiSW 0).
8. Maciej Brzeski, Przemysław Spurek. ***Uniform Cross-entropy Clustering***, Theoretical Conference of Machine Learning, Wydawnictwo UJ, Schedae Informaticae Journal, vol.25 (2017), ((pkt. MNiSW 11).
9. Maciej Brzeski. ***Schauder Bases in Banach Spaces***. Poster International Workshop for Young Mathematicians „Functional Analysis" (brak śladu publicznego) (2012), (pkt. MNiSW 0).

Liczba publikacji, cytowani i IF	Wartość
Liczba publikacji	Wartość
Według bazy DBLP	2
Według bazy Web of Science	4
Według bazy Scopus	5
Liczba cytowań ogółem	Wartość
Według bazy Web of Science	4 (2 bez autocytowań)
Według bazy Scopus	9 (6 bez autocytowań)
Indeks Hirsch	Wartość
Według bazy Web of Science	1
Według bazy Scopus	2

Ocena dorobku naukowego Doktoranta

Ważne publikacje, tematycznie związane dysertacją, to pozycja [1], [3], [4] i żadna z nich nie jest indeksowana w bazach bibliograficznych IT (*computer science bibliography)*. W dyscyplinie naukowej ***informatyka techniczna i telekomunikacja*** dorobek naukowy informatyków jest oceniamy wg liczby artykułów i publikacji, ich cytowań oraz indeksu Hirscha klasyfikowanych w 3 bazach bibliograficznych tj Digital Bibliography & Library Project (**DBLP**), Web of Science (**WoS**), **Scopus**. DBLP jest to najważniejsza światowa baza bibliograficzna dla informatyki (Computer Science). WoS to najbardziej prestiżowa i ekstremalnie selektywna światowa baza bibliometryczna – odpowiada za Journal Citation Reports (**JCR**) I Impact Factor (**IF**). Scopus to duża, komercyjna baza bibliometryczna firmy Elsevier, mniej selektywna niż WoS i obejmuje głównie nauki stosowane, informatykę, medycynę, ekonomię.
Zatem artykuł [1] z IF ((Impact Factor)= 0.7, *A Counterexample to the Order Relation Between Largest Common Subtree and Smallest Common Supertree*. opublikowany (w 2026), w **Annals of Combinatorics.** (pkt. MNiSW 100) nie jest indeksowany przez DBLP. Odpowiedź jest taka, że **Annals of Combinatorics** nie jest czasopismem informatycznym, a z definicji DBLP nie indeksuje matematyki. Cytowane w rozprawie artykuły zostały napisane starannie, precyzyjnym językiem, w związku tym nie wnoszę do nich uwag. Godny uwagi jest artykuł [1] w czasopiśmie z IF, dwie publikacje konferencyjne [5] z punktacją 70 pkt. MNiSW i [7] oraz trzy publikacje w czasopiśmie UJ, Schedae Informaticae Journal [4](20),[6](11),[8](11), z Doktorantem jako z pierwszym współautorem. W omówionych w dysertacji publikacjach zakładam, że głównym pomysłodawcą kluczowych rozwiązań i metod jest mgr Maciej Brzeski. Szkoda, że nie dołączono, potwierdzających powyższe, oświadczeń autorów publikacji i umieszczonych jako załączniki rozprawy – co byłoby pożądaną, dobrą praktyką. Osiągnięcia naukowe mgr Macieja Brzeskiego mieszczą się w zakresie wnioskowanej dziedziny nauki inżynieryjno-technicznych w dyscyplinie informatyka techniczna i telekomunikacja. Zebrane dane o dorobku naukowym doktoranta są wystarczające do wystawienia oceny pozytywnej.

7 **Ocena rozprawy doktorskiej**

Przedstawiona rozprawa została przygotowana w języku polskim, jako praca doktorska o charakterze wdrożeniowym. Rozprawa składa się z ośmiu rozdziałów, bibliografii, spisu tabel. Ogólna objętość rozprawy to 232 strony o spisie treści w formie wycinkowej dla ujęcia istotnej struktury pracy, jak niżej.

Ogólny stan wiedzy kandydata w zakresie Informatyki na podstawie oceny wybranych rozdziałów rozprawy (tj. 2, 5, 7**).**

Rozdział 2: *Zarządzanie danymi: od hurtowni po pochodzenie danych.* W rozdziale przedstawiono w sposób popularno-naukowy podstawy hurtowni danych, procesów ETL/ELT oraz zarządzania danymi (str. 10–21). W podrozdziale 2.3 opisano śledzenie

przepływu danych wyróżniając opatentowane, autorskie narzędzie MetaDex jako przykład niezależnego skanera, nie związanego z konkretnym katalogiem danych. W podrozdziale 2.4 opisano zagadnienie pochodzenia danych w kontekście zapewnienia ich wiarygodności, transparentności, audytowalności oraz zaufania do procesów analityczno-decyzyjnych.

Ocena rozdziału 2: Trywialno – techniczny opis zagadnień systemów hurtowni danych (DW) i systemów DSS i AdVD i niejednoznaczny kontekst, pobieżny wobec uwag niniejszej recenzji w punkcie 2 (Recstr. 3-7) pt. Cel rozprawy. Rozdział potwierdza średni stan wiedzy kandydata w zakresie Informatyki technicznej, w obrębie zagadnień Zaawansowanych Baz Danych i Hurtowni Danych. Natomiast opisany problem skutecznego wnioskowania o przepływie i pochodzeniu danych świadczy o dużej wiedzy praktycznej Doktoranta o rzeczywistych potrzebach minimalizowania ryzyka decyzyjnego przedsiębiorstw i instytucji;

Rozdział 5: *Podobieństwo procesów ETL-owych.* Celem rozdziału było opracowanie metody MetaDex/GED/Heurystyki zapewniającą formalną reprezentację grafów ETL i środowisko porównywanie procesów ETL w sposób spójny i skalowalny. Metoda ta pozwala uchwycić strukturę procesów ETL oraz zależności między poszczególnymi operacjami i ich wpływ na dane. Poprzez użycie mechanizmów unifikacji grafów oraz wstępnej klasteryzacji metoda ogranicza nadmiar porównań procesów ETL na poziomie pojedynczych operacji lub całych schematów przepływu danych. Przedstawiono również praktyczne aspekty użycia metody, tj. przygotowanie danych i dostosowania do efektywnego wykorzystania w rzeczywistych środowiskach. Metoda MetaDex/DAG-Edit może sugerując efektywne wzorce i unikania powielania błędów.

Ocena rozdziału 5: Wyniki uzyskane metodą MetaDex/DAG i ich szczegółową krytykę przedstawiono w niniejszej recenzji (Rec.str. 6) **Problem badawczy 1. *Analiza podobieństwa procesów ETL .... w kontekście zarządzania hurtownią danych.*** Doktorant wykazał się dużą wiedzy teoretyczną w obszarze podobieństwa drzew (rozdz.3) oraz podobieństwa grafów (rozdz. 4) tj. podał konkretną, formalnym definicję odległości edycyjnej GED dla grafów DAG, jako rzeczywiste uogólnienie odległości edycyjnej drzew. Doktorant przeprowadził dowód twierdzenia 3.45, gdzie wykazał brak podliniowej redukcji instancyjnej między odległością edycyjną a wyrównaniem drzew; nie tylko porządkuje istniejącą wiedzę, ale koryguje wcześniejsze ustalenia w literaturze. **Dobrze oceniam ten rozdział (str. 69 – 85) za jakość teoretyczną i wiedzę eksperymentalną Doktoranta w dyscyplinie Informatyka techniczna i telekomunikacja.**

Rozdział 7: *Wnioskowanie brakujących powiązań pochodzenia danych.* W rozdziale opracowano metodę wykorzystującą dwustopniowy model oparty o architekturę transformera (bi-koder i koder krzyżowy) oraz wyniki eksperymentalne uzyskane w środowisku rzeczywistym, gdzie zawodzą tradycyjne skanery kodu.

Ocena rozdziału 7: Doktorant rzetelnie opisuje metodykę postępowania z danymi, przyznając, że poleganie wyłącznie na metadanych (bez dostępu do wartości danych instancyjnych) jest istotnym ograniczeniem, wynikających z realiów bezpieczeństwa i prywatności przedsiębiorstw. Jednocześnie wykazuje, że mimo tych braków, zaproponowane rozwiązanie pozwala na skuteczne odtworzenie brakującej informacji o danych. Szczegółowy opis modelu dwustopniowego i jego krytykę znajduje się w niniejszej recenzji na Rec.str.7 i dla 11 jako **PROBLEM BADAWCZY 1 - JEGO CHARAKTER I ZNACZENIE Rozważmy realizację celu 1 i tezy 1 dotyczącej *problemu badawczego* 1. Analiza podobieństwa procesów ETL .... w kontekście zarządzania hurtownią danych.**

	Dobrze oceniam ten rozdział (str. 147 – 187) za jakość i wiedzę praktyczną i eksperymentalną doktoranta w dyscyplinie Informatyka techniczna i telekomunikacja. **Podsumowując ocenę rozprawy doktorskiej:** 1.) Doktorant wykazał się wiedzę badawczo-eksperymentalną w dyscyplinie Informatyka techniczna i telekomunikacja. 2.) Doktorant uzyskał wartościowe doświadczenie oparte o eksperymenty w zakresie *Podobieństwo procesów ETL-owych oraz Wnioskowania brakujących powiązań pochodzenia danych.* 3.) Wskazuje pewne praktyczne znaczenie wdrożonych rozwiązań metodę bazującą na dwupoziomowym modelu opartym o architekturę transformera (bi-koder i koder krzyżowy) i stwarzające potencjalne możliwości ich dalszego użycia. 4.) Rozbudowuje implementacyjnie kluczowo-ważną metodę MetaDex bazująca na koncepcji patentu Pub. No.: US 2012/0296862 współautorstwa Dawid Dudy - opiekuna pomocniczego Doktoranta. 5.) Doktorant wskazuje dalsze kierunki badań naukowych (rozdz. 7.6) tj. • Integracja statystyk danych np. profilowanie danych, co potencjalnie poprawiłoby skuteczność modelu. • Dostosowanie metadanych specyficznych dla danej firmy oraz dostosowaniu hiperparametrów modelu (np. progów podobieństwa). 6.) Doktorant wskazuje dalsze kierunki badań wdrożeń (rozdz. 8.2) nad zarządzaniem niejednorodnymi metadanymi poprzez rekonstrukcję grafów pochodzenia danych. 7.) ponadto: • stwierdzenia zawarte w rozprawie są godne zaufania, • dobra jakość prototypowego oprogramowania MetaDex/GED/Heurystyki, • rozprawa zawiera prawie kompletną bibliografią, • ogólnie rozprawa napisana poprawnym i zrozumiałym językiem polskim z zachowaniem staranności. **Uwagi krytyczne** 1. Doktorant uważa, że możliwości modeli generatywnych wydają się obecnie prawie nie ograniczone, a badania naukowe nad nimi fascynujące. Wg mnie przyszłością sztucznej inteligencji są modele Generatywnych Przestrzeni Danych i Algorytmów (modele GPDiA). Modele GPDiA uczą się, jak dane, algorytmy i architektury PDiA są dystrybuowane, poszukują architektur specjalizowanych, schematów i wzorców, a następnie generują nowe architektury IT korzystając z tej wiedzy. Wstępną wiedzę na temat modeli GPDiA przedstawiłem w punkcie 3 niniejszej recenzji (Rec.str. 12). **2.** Doktorant w kolejnym kroku badań mógłby zaimplementować przykładową **Nowa Architekturę Modelu MetaDex/GDE/2EnKodery** (Rec.str. 17) lub podobną. 3. Doktorant nie wykazał współautorstwa posiadania patentów, wniosków racjonalizatorskich, czy też zgłoszonych wniosków patentowych. 4. Brak oświadczeń o udziale we współautorstwie publikacji. 5. Brak spisu nagród i wyróżnień autora.
8	**Inne uwagi**[1] Zawartość merytoryczna rozprawy jest spójna z metodyką badawczą inżynierii systemów pozwalającą na realizację założeń i celów badawczych. Autor w sposób sprawny przedstawia cały przeprowadzony cykl badawczy, jak również przejrzyście prezentuje wnioski z eksperymentów. Zatem stwierdzam, że oceniana rozprawa

	doktorska prezentuje umiejętność samodzielnego prowadzenia pracy naukowej przez kandydata do stopnia doktora. Oceniając rozprawę pragnę podkreślić, iż została ona wykonana na dobrym poziomie wdrożeniowym o dużym znaczeniu praktycznym i jest wartościowa z punktu widzenia pogłębienia wiedzy w dyscyplinie Informatyka Techniczna i Telekomunikacja. Wnosi ona także oryginalny wkład eksperymentalny.
9	**Podsumowanie** Biorąc pod uwagę opinie zaprezentowane w poprzednich punktach i wymagania zdefiniowane przez art. 187 Ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (z późniejszymi zmianami) [1] moja ocena rozprawy pod względem trzech podstawowych kryteriów jest następująca: A. Czy rozprawa zawiera oryginalne rozwiązanie problemu naukowego? (wybierz jedną opcję stawiając znak X) [X] *Zdecydowanie TAK* [ ] *Raczej TAK* [ ] *Trudno powiedzieć* [ ] *Raczej NIE* [ ] *Zdecydo NIE* B. Czy po przeczytaniu rozprawy zgadzasz się, że kandydat posiada ogólną wiedzę teoretyczną w dyscyplinie Informatyka techniczna i telekomunikacja? [ ] *Zdecydowanie TAK* [X] *Raczej TAK* [ ] *Trudno powiedzieć* [ ] *Raczej NIE* [ ] *Zdecydowanie NIE* C. Czy kandydat posiada umiejętność samodzielnego prowadzenia pracy naukowej? [X] *Zdecydowanie TAK* [ ] *Raczej TAK* [ ] *Trudno powiedzieć* [ ] *Raczej NIE* [ ] *Zdecydo NIE* Po zapoznaniu się z pracą doktorską ***mgr Macieja Brzeskiego*** zatytułowanej:. ***„Analiza podobieństwa procesów ETL i pochodzenia danych w kontekście zarządzania hurtownią danych”*** stwierdzam, że: **przedstawiona rozprawa spełnia**

[1] http://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20190000276

	właściwe ustawowe wymagania stawiane rozprawom doktorskim oraz mieści się w dyscyplinie Informatyka Techniczna i Telekomunikacja i w związku z powyższym **wnioskuję o przyjęcie rozprawy doktorskiej oraz dopuszczenie** **mgr Macieja Brzeskiego do publicznej obrony w wyżej wspomnianej dyscyplinie.** M. Brzeski (podpis Recenzenta)