Р Е Ц Е Н З И Я

от *проф. д-р Боян Паскалев Бончев*
СУ „Св. Климент Охридски”, ФМИ, катедра „Софтуерни технологии”

на дисертационен труд на тема „Методи за реализация на софтуерни системи за обработка на големи данни” (Methods for implementation of data-intensive software systems)

с автор *Симеон Стоичков Емануилов* – редовен докторант
към катедра „Софтуерни технологии”, ФМИ-СУ,
за придобиване на образователна и научна степен *„Доктор“*
в професионално направление 4.6 „Информатика и компютърни науки“,
докторска програма „Софтуерни технологии” – Софтуерно инженерство

Този документ представлява рецензия на дисертационен труд, научни публикации и автореферати на английски и български език за придобиване на образователна и научна степен „доктор“ в професионално направление 4.6 „Информатика и компютърни науки“, докторска програма „Софтуерно инженерство“, изготвена съгласно Заповед №РД-38-283/09.06.2025г. на Ректора на Софийския университет и Протокол №1 от 12.06.2025г. от първото заседание на научното жури.

1. Актуалност на проблема

С увеличаването на обемите данни, създаването на нови архитектурни методи, техники за представяне на данни и стратегии за обработка става от съществено значение. Напредъкът в тази област е жизненоважен за управлението на обработката, транспорта и съхранението съвременните големи данни и за реализирането на пълния потенциал на технологиите, базирани на големи данни, в сектори като здравеопазване, финанси и научни изследвания.

Представеният от Симеон Емануилов дисертационен труд е в актуална и динамично развиваща се област, свързана с проектирането на софтуерни системи с интензивна обработка на данни. Дисертацията е фокусирана върху предизвикателствата при проектирането на софтуерни системи с интензивна обработка на данни, изследвайки решения за подобряване на ефективността, мащабируемостта и интерпретируемостта в процесите на обработка на големи данни. Изследването обхваща разработването, оптимизирането и оценяването на иновационни методи за реализиране на софтуерни системи с интензивна обработка на данни. По този начин трудът допринася за ефективното използване на големи данни със значими социални и икономически ползи.

2. Познаване на състоянието на проблема

Дисертационният труд и представените публикации показват ясно, че докторантът е отлично запознат с разглежданата проблематика. Цитирани са голям брой литературни източници - общо 152, като на четири от тях докторантът е съавтор. Показателно за актуалността на изследвания проблем е, че всички цитирани

литературни източници са на английски език, като повечето от тях са съвременни (от 2015 г. насам), като само три литературни източника са от 90-те години на миналия век.

3. Обща характеристика на представения дисертационен труд

Дисертационният труд е написан на английски език и се състои от речник на термините, въведение, изложение в шест глави, заключение, обобщение на синергетичната интеграция на използваните в изследването подходи, справки за научните приноси, публикации и презентации на докторанта, последствия за софтуерните системи с интензивно използване на данни, насоки за бъдеща работа, списък на използваната литература (библиография), четири приложения и декларация за оригиналност, представени на общо *157 страници*. Включени са 21 фигури, 19 таблици и 9 извадки от програмен код, като е и приложена декларация за оригиналност. Съдържанието и структурирането му се определят от поставените задачи и следването на избраната методология.

Обект на изследване са софтуерни системи, интензивни на данни, и методите за тяхната реализация. Това включва системи, които управляват и анализират големи обеми данни, потенциално скалируеми до милиарди записи, което създава уникални предизвикателства по отношение на производителността и мащабируемостта. *Предметът на изследването* е разработването, оптимизирането и оценяването на нови методи и техники за реализиране на софтуерни системи с интензивна обработка на данни. Оттук и *целта на дисертационния труд* е да подобри реализацията на софтуерни системи с интензивна обработка на данни, чрез създаване и оценка на нови методи за управление и анализ на големи колекции от данни. Целта включва идентифициране на ограниченията на текущите подходи и предлагане на решения, които подобряват мащабируемостта, производителността и интерпретируемостта. Изследването акцентира върху оптимизацията на представянето на данни, съхранението, индексирането и техниките за извличане, като се фокусира върху базите данни, ориентирани към колонен формат, и алгоритми за търсене на сходство върху набори от данни с милиарди записи. Освен това, то цели да направи сложните структури от данни по-достъпни, като се трансформират високоизмерни вектори в интерпретируеми форми.

За постигане на поставената в дисертационния труд цел са дефинирани *5 основни задачи*, които могат да се обобщят така:

1. Изследване на архитектурни подходи за системи с интензивна обработка на данни, с оценка на тяхната ефективност относно мащабируемост и производителност.

2. Разработка на усъвършенствани техники за индексиране и търсене на многомерни данни в набори от данни с милиарди записи, с фокус върху решения за търсене на сходство.

3. Изследване на методи за подобряване на интерпретируемостта на сложни представяния на данни, особено трансформирането на високоизмерни числови вектори в достъпни формати.

4. Изследване на оптимизирани модели на данни за ефективно съхранение, извличане и обработка в среди с интензивна обработка на данни с висока скорост на приемане.

5. Провеждане на експериментални изследвания и сравнителни анализи за оценка на ефективността на предложените подходи в реални сценарии, включително търсене на сходство в голям мащаб и интерпретация на данни.

При разработване на дисертационния труд е приложен систематичен подход, съответстващ на поставените задачи, като е избрана подходяща методология за провеждане на научното изследване. Глава 1 представя преглед на системите с интензивна обработка на данни и свързаните с тях предизвикателства. Тя разглежда разликите между системите, базирани на машинно обучение, и традиционните софтуерни системи, акцентирайки на адаптивната им природа и значението на данните в съвременната разработка на софтуер. Обсъждат се основни концепции като високоизмерни данни и търсене на сходство, както и техники за намаляване на размерността и различни архитектурни стилове за системи с интензивна обработка на данни.

Глава 2 се фокусира върху техники за индексиране и клъстеризиране, необходими за управление и търсене в големи набори от данни. Тя установява предизвикателствата, свързани с търсенето на сходство при високоизмерни данни, особено когато наборите от данни достигат милиарди записи, при съществуващи индексиращи подходи като инвертиран файл (IVF), квантуване на продукта (PQ) и графово-базирани методи.

Глава 3 представя нов хибриден алгоритъм за индексиране, който адресира предизвикателствата на търсене на сходство в мащаб от милиард. Предложението включва интеграция на плътни векторни вложения и дискретни филтриращи атрибути в единна индексна структура, използваща усъвършенстван IVF-Flat индекс. Главата описва стъпките за изграждане на хибридни вектори и подчертава динамичната стратегия за управление на паметта, която осигурява ефективно справяне с набори от данни, надвишаващи наличната оперативна памет. В допълнение е демонстрирана ефективността на предложената хибридна индексираща стратегия.

Глава 4 се занимава с подобряване на интерпретируемостта на плътните вектори и представя нова техника, наречена LangVec, която може да подобри разбирането на векторните данни. Основният принос е базираното на персентили съпоставяне на векторни величини към думи от предварително дефиниран лексикон. Главата подробно описва тристепенния процес на дефиниране на лексикон, изчисление на персентилите и съпоставянето на вектори към думи. Резултатите от бенчмарк тестовете показват ефективността и мащабируемостта на подхода, използвайки както синтетични, така и реални набори от данни.

Глава 5 разглежда колоно-ориентирани модели на данни за системи с интензивна обработка на данни и предлага денормализиран, колоно-ориентиран модел, оптимизиран за уеб системи. Основният акцент е поставен върху операциите SELECT, INSERT и аналитичните заявки върху исторически данни. Главата предлага алгоритъм за динамично генериране на колонната структура на базата на предназначението на полетата и разглежда оптимизационни техники като компресия на данни и организиране на данните по физически критерии. Проведен е сравнителен анализ на производителността спрямо традиционна редово-ориентирана база данни, който демонстрира подобрения в производителността за различни типове заявки.

Глава 6 представя резултати от експериментални изследвания и сравнителни анализи за оценка на ефективността, ефикасността и мащабируемостта на предложените методи за внедряване на софтуерни системи с интензивно използване на данни. Включва проучване за търсене на сходство в мащаб от милиард с набора от данни LAION-5B, което показва значително намаление на времето за търсене чрез паралелна обработка. Главата също така описва интеграция на предложените методи в производствена семантична търсачка, демонстрирайки по-кратки времена за отговор при търсене на сходство и подобрена производителност в аналитични задачи. Включени са и резултати от тестове с отворен набор от данни, което предоставя допълнителни доказателства за ефективността на представените методи.

Заключението обобщава основните находки, синтезира резултатите от всички глави и обсъжда последиците за софтуерните системи с интензивна обработка на данни. То предлага насоки за бъдещи проучвания и разработки, като акцентира на потенциални приложения на представените методи.

Като цяло дисертационният труд е балансиран, въпросите са изложени в логическа последователност и обвързаност, стилът на изложението е научен. Текстът е добре структуриран, при което основните акценти, изводи и резултати са откроени, което подпомага обобщаването на научно-приложните и приложните приноси.

4. Оценка на получените резултати

В резултат на реализираното изследване на текущото състояние в предметната област са *установени научно-приложни проблеми* за необходимостта от проектиране и валидиране на софтуерни системи с интензивна обработка на данни. Получените научни и научно-приложни резултати от работата на дисертанта са съответствие с дефинираната цел и поставените задачи. *Методиката за разработване на дисертационния труд* включва прилагане на подходи за моделиране и на методи за анализ и синтез, както и извършване на експерименти и оценъчен анализ на адекватността на предложените решения. Използван е интегриран подход, валидиран чрез обширни измервания на производителността и практическо внедряване, който показва как методологията и инструменти работят заедно, за да създадат по-ефективна, интерпретируема и мащабируема архитектура за системи с интензивна обработка на данни. Синергията между компонентите надхвърля простото комбиниране, създавайки комплексно решение, което има практическа приложимост.

Положителна оценка заслужава значителният обем от цитирани литературни източници и нейната прецизна интерпретация. Идентифицирани са предизвикателствата за проектирането, създаването и управлението на софтуерни системи с интензивна обработка на данни. В резултат на проучването докторантът представя изводи, водещи до създаването интегрирана рамка, в която множество компоненти работят заедно за решаване на установените предизвикателства в системите с интензивни данни. Комбинацията от хибриден индекс, техника за интерпретация LangVec и колоно-образен модел на данни се използва за създаване на цялостно решение за ефективно управление и анализ на големи обеми от данни. Дизайнът на хибридния индекс е оптимално съчетан с колоно- ориентирания модел на съхранение, което позволява ефективно управление на плътни вектори и свързаните с тях атрибути. Измерванията на производителността показват, че системата осигурява времена под секунда за запитвания за набори от данни, надвишаващи 30 милиона вектора. Интеграцията на LangVec с хибридната структура на индекса създава мощна система за бързо извличане и интерпретация на резултатите. Инструментът gotake допълнително подобрява възможностите за придобиване на данни, постигайки значителни увеличения на скоростта на изтегляне. Колоно-ориентираният модел е особено ефективен при обработката на данни с висока скорост, което допълва изискванията за обработка на вектори. Освобождаването на обработения COCO набор от данни с вектори от CLIP ViT-L/14 представлява значителен принос за оценка на системите за търсене на сходства. Практическата реализация в Similarix убедително демонстрира как комбинираното използване на тези технологии води до ефективно търсене на сходства. Подходът на хибридно индексиране в милиарден мащаб показва ефективност, близка до линейната на мащабиране при по-голям брой възли.

5. Оценка на приносите

Приносите на дисертационния труд могат да бъдат обобщени в две групи както следва:

А. Научно-приложни приноси

1. Качествен анализ на ключови характеристики на архитектурни подходи и стилове за преодоляване на предизвикателствата в системите за обработка на интензивни данни, с оценка на приложимостта им за подобряване на скалируемостта и производителността.
2. Разработка на нови методологии:
    a. за индексиране за търсене на сходства в големи масиви от данни, включваща метод за комбиниране на вектори с филтриращи атрибути, подобрена индексна структура с многомерно филтриране, и ефективна стратегия за управление на паметта за обработка на набори от данни.
    b. за интерпретиране на високо-измерни вектори с нова техника за картографиране, трансформиране на величините на векторите в лексикални маркери, адаптивен алгоритъм за разпределение на лексикона за балансирано представяне и метод за обработка на вектори в блокове (наречена LangVec).
3. Разработка на ориентиран към колони модел на данни, оптимизиран за системи с интензивни данни – с денормализирана схема, алгоритъм за динамично генериране на колони и техники за оптимизация на ефективността на съхранение, скоростта на извличане и възможностите за обработка.

Б. Приложни приноси

1. Реализация и внедряване на софтуерна система, реализираща методологията за интерпретация на вектори; семантична търсачка с хибриден индексен подход; реализацията на модела на данни и допълнителни инструменти за повишаване на общата ефективност на системата.
2. Разработка на рамки за оценка и практическа валидация на производителността и интерпретируемостта на LangVec чрез тестове с набори от данни в реални сценарии.

Изследването, представено в тази дисертация, демонстрира последователен подход към решаването на многостранните предизвикателства на системите с интензивни данни. Архитектурният анализ предоставя основа за проектиране на системи, върху която хибридният индексен подход и колоно-ориентираният модел на данни изграждат ефективни решения за управление и извличане на данни. LangVec подобрява интерпретируемостта на сложните представяния на данни и свързва разликата между машинната ефективност и човешкото разбиране.

6. Оценка на публикациите по дисертационния труд

По тематиката на дисертационния труд са докторантът представя три научни публикации на английски език, от които две са в научни списания и една е в сборник на международна конференция, индексиран в IEEE Xplore Digital Library. Всичките публикации са в съавторство с научния ръководител, като и в трите докторантът е първи съавтор. Двете публикации в списания са в издания с импакт фактор (IF=1.1/Q4(Comp. Science, Inf. Sys.)/2024 и IF=0.3/Q4(Education & Educ. Research)/2024), като едното има и импакт ранг (SJR=0.358/Q2(Computer Science)/2024).

7. Качества на автореферата

Авторефератът е изготвен в обем от 53 машинописни страници на български език и и 46 страници на английски език, като отговаря на общоприетите изисквания. Той отразява пълно, коректно и адекватно постигнатите в дисертационния труд резултати и заявените от докторанта приноси.

8. Критични бележки, препоръки и въпроси

Дисертационният труд е оформен много старателно и грижливо, като всичките ми забележки и препоръки относно съдържанието и презентирането му от двете предварителни представяния на труда са отразени в настоящата му версия. Така например практическите приноси са допълнени с препратки към документи, показващи тяхното приложение в реални случаи на употреба (като научно-изследователски проекти, практически софтуерни модели, индустриални софтуерни решения); добавена е таблица, която указва за всеки претендиран от докторанта принос в коя глава от труда е представен, коя от публикациите го описва, и къде/как е използван в реални сценарии, и др. (особено за приложните приноси).

Трябва да се отбележи, че две от статиите на докторанта са цитирани три пъти от различни чуждестранни учени.

9. Заключение

Общата ми оценка за дисертационният труд и научните публикации на Симеон Стоичков Емануилов е **положителна**. Предвид оригиналните научно-приложни и приложни приноси, както и постигнатите значими практически резултати, считам, че докторантът притежава задълбочени теоретични знания по съответната област и способности за самостоятелни научни изследвания. Дисертационният труд отговаря на изискванията на Закона за развитието на академичния състав в Република България (ЗРАСРБ), Правилника за прилагане на ЗРАСРБ и Правилника за условията и реда за придобиване на научни степени и заемане на академични длъжности в СУ „Св. Климент Охридски". Това ми дава основание да предложа на уважаемото научно жури да присъди на Симеон Стоичков Емануилов образователната и научна степен „Доктор" в професионално направление 4.6 „Информатика и компютърни науки", докторска програма „Софтуерни технологии" – Софтуерно инженерство.

24.08.2025 г. Подпис: ……

гр. София /Боян Бончев/