Министерство образования Республики Беларусь

Учреждение образования

«Белорусский государственный университет

информатики и радиоэлектроники»

УДК

ПОПЦОВ

Алексей Андреевич

МИГРАЦИЯ СЛОЖНОСТРУКТУРИРОВАННЫХ МОДЕЛЕЙ ДАННЫХ ИЗ РЕЛЯЦИОННОЙ БАЗЫ ДАННЫХ В ГРАФОВУЮ

АВТОРЕФЕРАТ

диссертации на соискание
степени магистра информатики и вычислительной техники

по специальности 1-40 81 03– Искусственный интеллект

Минск 2017

Работа выполнена на кафедре искусственного интеллекта учреждения образования «Белорусский государственный университет информатики и радиоэлектроники»

Научный руководитель: **РОМАНОВ Владимир Ильич,**
кандидат технических наук, доцент кафедры биомедицинской информатики учреждения образования «Белорусский государственный» ФПМИ

Рецензент: **ПОТТОСИНА Светлана Анатольевна,**
кандидат физико-математических наук, доцент, доцент кафедры экономической информатики учреждения образования «Белорусский государственный университет информатики и радиоэлектроники»

Защита диссертации состоится «27» июля 2017 г. года в $9^{\underline{00}}$ часов на заседании Государственной комиссии по защите магистерских диссертаций в учреждении образования «Белорусский государственный университет информатики и радиоэлектроники» по адресу: 220013, г.Минск, ул. Платонова, 39, 5 уч.корп., ауд. 607, тел.: 293-80-92, e-mail: kafiit@bsuir.by.

С диссертацией можно ознакомиться в библиотеке учреждения образования «Белорусский государственный университет информатики и радиоэлектроники».

ВВЕДЕНИЕ

Нынешний век – век информационных технологий. Через современные информационные системы проходят огромные массивы данных. При таких объемах данных проектирование и разработка подсистем хранения информации в программном обеспечении становится непростой задачей для инженеров. Кроме того стоит, учитывать, что мы живем во взаимосвязанном мире, где важны не только свойства предмета, но и его взаимосвязи с другими предметами, а значит эти взаимосвязи должны быть отражены в моделях хранения данных. В течение нескольких десятилетий инженеры пытались приспособить реляционные базы данных для работы с естественными наборами данными. Но, так как реляционные базы данных изначально предназначены для обработки таблиц, попытки моделирования взаимосвязей реального мира являются затруднительны. Реляционные базы данных включают механизмы взаимосвязей, но применяется он только на этапе моделирования, как средство объединения таблиц и не всегда в полной мере способны удовлетворить потребности инженеров т.к. зачастую требуется устранить неоднозначность семантики взаимосвязей, связывающих объекты, а также квалифицировать их вес и силу. Кроме того, рост количества связей приводит в реляционном мире к увеличению количества операций соединений, которые снижают производительность системы, что будет в дальнейшем показано в данной работе. Использование графовых структур для хранения информации способно в какой-то мере помочь в решении обозначенные выше проблемы

В упрощенном варианте графовую структуру можно определить как набор вершин и взаимосвязей. В графах объекты представлены узлами, а способы, которыми эти объекты соединены между собой – взаимосвязями. Эта универсальная и выразительная структура позволяет моделировать всевозможные сценарии, от постройки космической ракеты до строительства системы дорог. Графы чрезвычайно полезны при анализе самых разных наборов данных в таких областях, как наука, государственное управление и бизнес.

Так для хранения графовых структур были разработаны системы управления графовыми базами данных или графовые базы данных, которые поддерживают методы создания, чтения, изменения и удаления основанных на графовых моделях данных. Решение о хранение информации о связях в подобных системах является естественным т.к.

упрощает построение моделей. Несмотря на это, мы живем в прагматичном мире бюджетов, корпоративных стандартов и коммерческих правил. Предоставляемый графовыми базами данных новый метод моделирования данных сам по себе не является достаточным основанием замены устоявшихся и понятных платформ обработки данных. По этой причине необходимо проводить тщательный анализ выгод и потенциальных проблем, которые может принести использование графовой СУБД.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования

Существует множество книги и статьи в полной мере описывают достоинства и недостатки использования графовых моделей при проектировании подсистемы хранения данных, кроме того они предоставляют практические руководства об использовании конкретных СУБД. Однако процесс перехода с существующей реляционной модели на графовую описан недостаточно. Нет критериев необходимости такого перехода и паттернов внедрения графовой БД в уже работающий программный продукт на реляционной БД.

Цель и задачи исследования

Цель данной работы провести миграцию реляционной модели в графовую, выявив паттерны упрощения подобного процесса в дальнейшем. В рамках данной цели было поставлено несколько задач:

- Разобраться в различиях реляционной и графовой модели
- Произвести миграцию существующего приложения с реляционной модели хранения на графовую
- Разработать паттерны миграции с реляционной модели на графовую

Область исследования

Содержание диссертационной работы соответствует образовательному стандарту высшего образования второй ступени (магистратуры) специальности 1-40 81 03 «Искусственный интеллект».

Теоретическая и методологическая основа исследования

В основу диссертации легли изыскания зарубежных и отечественных ученых в области хранения данных в виде графовых моделей.

В качестве инструментальных средств использовались объектно-ориентированный язык программирования *Java*, фреймворк *Spring, база данных Neo4j*.

Научная новизна

Научная новизна и значимость диссертации заключается в разработке подхода к миграции существующей реляционной модели данных на графовую, который мог бы быть многократно использован в других системах.

Теоретическая значимость диссертации заключается в описании различий между графовой моделью данных и реляционной.

Практическая значимость результатов исследования заключается в сравнении производительности графовой БД и реляционной БД .

Основные положения, выносимые на защиту

1. Обоснование того факта что существующие источники информации о графовых базах данных в недостаточной степени описывают процесс миграции с реляционной модели данных на графовую модель,.
2. Алгоритм миграции упрощающий процесс переход с реляционной базы данных на графовую базу данных. В основе алгоритма лежит различие в проектировании реляционной модели и графовой, так таблицы сущности трансформируются в узлы графовой структуры, а таблицы взаимосвязей в ребра.
3. Результат тестирования позволяющий доказать факт, того что использование нескольких баз данных для системы хранения в ряде случаев предоставляет возможность нивелировать недостатки каждой базы данных по отдельности

Апробация диссертации и информация об использовании ее результатов

Основные теоретические результаты и законченные этапы диссертационной работы были изложены в рамках внутренней конференции компании Graphaware.

Публикации

Изложенные в диссертации основные положения и выводы были использованы в официальном руководстве Neo4j https://neo4j.com/developer/graph-db-vs-rdbms .

Структура и объем работы

Диссертация состоит из введения, общей характеристики работы, двух глав с краткими выводами по каждой главе, заключения, библиографического списка и приложений.

Во введении рассмотрены общие проблемы реляционных БД и то каким образом использование графовых моделей может помочь их решить.

В первой главе рассмотрена существующая литература, связанная с темой диссертации, а также сделано краткое сравнение существующих графовых БД.

Вторая глава посвящена практической части миграции с реляционной БД на графовую. Происходит обзор существующей реляционной модели, вырабатывается алгоритм перехода на графовую модель и сам переход. В качестве итогов проводится тестирование производительности полученных результатов.

В приложении представлен значимый исходный код для понимания диссертации.

Общий объем диссертационной работы составляет 48 страниц. Из них 35 страниц основного текста, 14 иллюстраций, 3 таблиц, библиографический список из 13 наименований, 3 приложения.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** рассмотрены общие проблемы реляционных БД и то каким образом использование графовых моделей может помочь их решить.

В **общей характеристике работы** показана актуальность проводимых исследований, степень разработанности проблемы, сформулированы цель и задачи диссертации.

В первой главе сделан обзор существующих источников информации на тему магистерской диссертации. Так, дан обзор 5, а также 5 статьям зарубежных авторов. Кроме того, произведен обзор трех существующих графовых баз данных, сделано сравнение способов хранения графовой модели в каждой из СУБД.

Вторая глава, состоящая и раздела проектирования и реализации, посвящена практической части диссертации. Изначально дается подробное описание реляционной модели, которая будет использована в качестве тестовой для проведения миграции. Озвучиваются существующие проблемы модели: недопустимо низкая скорость выполнения запросов, требующих операции JOIN и плохая читабельность существующих запросов. Далее дано описание способов построения графовых моделей. Для дальнейшей работы выбирается графовой модели со свойствами и метками. После этого выстраивается графовая модель, способная, с точки зрения функциональности не только заменить существующую, но и решить оговоренные выше проблемы. После того как модель построена, полученный опыт обобщается в алгоритм действий для ускорения повторной миграции в других системах.

В заключительной части происходит сравнение производительности (одна из основных проблем существовавшей модели) старой системы с новой. Результатом данного сравнения являются таблицы:

16 000 вершин и 600 000 ребер				
Тестовая БД	Время, с			Размер БД
	Запрос А	Запрос Б	Запрос В	
Neo4j	0.2	0.3	0.3	60
Mysql	0.1	0.7	1	45

100 000 вершин и 1 200 000 ребер				
Тестовая БД	Время, с			Размер БД
	Запрос А	Запрос Б	Запрос В	
Neo4J	0.3	1.1	1.3	3023
Mysql	0.2	4	6	2835

4 000 000 вершин и 38 000 000 ребер				
Тестовая БД	Время, с			Размер БД
	Запрос А	Запрос Б	Запрос В	
Neo4J	0.7	6	8	13867
Mysql	0.5	300	849	10586

10 000 000 вершин и 90 000 000 ребер				
Тестовая БД	Время, с			Размер БД Мб
	Запрос А	Запрос Б	Запрос В	
Neo4J	1.5	14	17	31004
Mysql	1.2	1398	7392	26209

Из результатов нагрузочного тестирования видно что Neo4j в большинстве запросов превосходит в скорости выполнения Mysql на 1-2 порядка. Это связано с большим количеством операций JOIN в запросах Mysql. В то же время размер занимаемого места на жестком диске у Mysql примерно в 1.5 раза меньше. Однако считаю, что если бизнес задача требует от ПО строить аналитику по большому количеству сущностей, которая требует выполнения операций слияния, то выигрыш в производительности играет более существенную роль чем размер БД на жестком диске. Стоит также отметить, что при количестве операций слияния меньше 5 реляционная база данных показывает лучшую производительность. По этой причине, в случае, зачастую оправдано использование нескольких баз данных в одной системе.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Выполнено сравнение реляционной модели данных с графовой [2];

2. Разработан алгоритм миграции с реляционной модели в графовую модель, который в дальнейшем был использован в документации Neo4j [1];

3. Произведено тестирование работы системы на реляционной БД и на графовой;

Рекомендации по практическому использованию результатов

Полученные результаты частично внедрены в руководство к базе данных Neo4j.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ

1. Алексей Попцов в составе «Neo4j Staff From Relational to Neo4j» https://neo4j.com/developer/graph-db-vs-rdbms/.

2. Алексей Попцов в составе «Neo4j Staff The Database Model Showdown: An RDBMS vs. Graph Comparison» https://neo4j.com/blog/database-model-comparison/.

РЭЗЮМЭ

Папцоў Аляксей Андрэевіч

Міграцыя

складана структураваных мадэляў дадзеных

з рэляцыйнай баз дадзеных у графавыя

Ключавыя словы: база дадзеных, графы.

Мэта працы: распрацоўка абагульненага алгарытму міграцыі дадзеных з рэляцыйнай БД у графавая БД.

Атрыманыя вынікі і іхнавізна: Выраблена параўнанне рэляцыйнай мадэлі дадзеных з графавымі. Распрацаваны спіс інструкцый для творы міграцыі модуля захоўвання дадзеных у сістэмы, якая выкарыстоўвае рэляцыйную БД на модуль выкарыстоўвае камбінацыю рэляцыйнай і графавай БД. Выпрацаваны крытэрыі абгрунтаванасці падобнай міграцыі.

Ступень выкарыстання: вынікі часткова выкарыстаны ў кіраўніцтве да БД Neo4j.

Вобласць ужывання: распрацоўка праграмнага забеспячэння.

РЕЗЮМЕ

Попцов Алексей Андреевич

Миграция сложноструктурированных моделей данных из реляционной баз данных в графовые

***Ключевые слова*:** база данных, графы.

***Цель работы*:** разработка обобщенного алгоритма миграции данных из реляционной БД в графовую БД.

***Полученные результаты и их новизна*:** Произведено сравнение реляционной модели данных с графовой. Разработан список инструкций для произведения миграции модуля хранения данных в системы, использующей реляционную БД на модуль использующий комбинацию реляционной и графовой БД. Выработаны критерии обоснованности подобной миграции

***Степень использования*:** результаты частично использованы в руководстве к БД Neo4j.

***Область применения*:** разработка программного обеспечения.

SUMMARY

Poptsov Alexey Andreevich

Migration of complex data models

from a relational database to a graph database

***Keywords*:** database, graphs.

***The object of study*:** to develop a generalized algorithm for migration data from a relational database to a graph database.

***The results and novelty*:** The comparison of the relational data model with the graph model was done. A list of reusable instructions to do migration of data models from a relational database to a graph database for developers was made.

***Degree of use*:** the results are partially used in the manual for the Neo4j database.

Sphere of application: software development.