Санкт-Петербургский государственный университет

Математическое обеспечение и администрирование информационных систем

Группа 22.М04-мм

Модификация процессов сбора, хранения и управления данными для апробации алгоритмов анализа графов

Соболь Дарья Валерьевна

Отчёт по учебной практике
в форме «Производственное задание»

Научный руководитель:
старший преподаватель, к.ф-м.н., Р.Ш. Азимов

Санкт-Петербург
2024

Оглавление

Введение

Представление данных при помощи помеченных графов широко используется в таких областях, как биоинформатика, статический анализ кода, а также в других сферах. При работе с такими данными часто возникают запросы навигации и поиска путей, удовлетворяющих определенным ограничениям. Результаты обработки таких запросов обычно представляют собой набор отношений между вершинами графа. Один из естественных способов определить эти отношения — указать соответствующие пути, используя формальные грамматики над алфавитом меток ребер, которые могут быть выражены с помощью контекстно-свободных грамматик. Возникает вопрос о необходимости разработки и реализации алгоритмов поиска путей с контекстно-свободными ограничениями (CFPQ). Учитывая широкое применение контекстно-свободных запросов в перечисленных выше областях, критически важной становится потребность в измерении производительности алгоритмов, реализующих эти запросы. Для демонстрации применимости алгоритма на практике требуется проведение экспериментального исследования на помеченных графах, соответствующих реальным данным. Однако поиск и подготовка таких графов весьма сложны и могут занять значительное время. Модернизация хранилища данных для такого проекта контекстно-свободных запросов к помеченным графам может принести несколько значительных преимуществ. Новая система хранения данных может быть оптимизирована для быстрого выполнения запросов навигации и поиска путей в помеченных графах, что повысит производительность алгоритмов CFPQ. Также, модернизированное хранилище данных может обеспечить более эффективную подготовку реальных помеченных графов для экспериментального исследования, что упростит проведение тестов и экспериментов. В результате, модернизация хранилища данных может способствовать повышению точности и скорости анализа данных, необходимого для разработки и оптимизации алгоритмов контекстно-свободных запросов к помеченным графам.

1. Постановка задачи

Целью данной работы является модификация процессов сбора, хранения и управления данными для апробации алгоритмов анализа графов в проекте CPFQ_Data. Для достижения поставленной цели были выделены следующие за- дачи.:

1. Провести обзор конкурентных продуктов;
2. Проанализировать и сравнить существующие подходы к хранилищам данных:
3. Изучить способы модернизировать проект CPFQ_Data.

2. Обзор

2.1. Хранение данных

В настоящее время в области хранилищ данных приняты различные стандарты, целью которых является установление единой методологии и формата для организации и обработки данных. Одним из таких стандартов является SQL (Structured Query Language), который широко используется для управления реляционными базами данных. SQL предоставляет стандартизированный набор команд для создания, изменения и управления данными, а также для выполнения запросов к базам данных.

Для хранения и обработки больших объемов данных активно применяются технологии NoSQL (Not Only SQL). NoSQL базы данных позволяют работать с разнообразными типами данных, не ограничиваясь схемой реляционных баз данных. В этой области существует ряд стандартов и принципов, таких как ACID (Atomicity, Consistency, Isolation, Durability) и CAP (Consistency, Availability, Partition tolerance), которые определяют требования к стабильности, доступности и распределенности данных [3].

Для области анализа больших данных (Big Data) широко применяются стандарты и технологии, такие как Apache Hadoop и Apache Spark, которые предоставляют средства для обработки и анализа данных в распределенной среде. Также следует отметить стандарты для представления и передачи данных, такие как JSON (JavaScript Object Notation) и XML (eXtensible Markup Language), которые используются для обмена структурированными данными между приложениями.

Таким образом, в современной области хранилищ данных существует множество стандартов и технологий, каждый из которых предназначен для определенных задач и обеспечивает эффективное управление и обработку информации в соответствии с требованиями современных бизнес-процессов и технологических трендов.

2.2. Обзор аналогов

Цель данного обзора состоит в том, чтобы провести обзор существующих аналогичных проектов, исследований или продуктов, связанных с темой настоящей работы. Анализ аналогов позволит выявить особенности и преимущества существующих подходов, а также определить их недостатки и возможности для улучшения.

1. **Matrix Market & The SuiteSparse Matrix Collection (formerly the University of Florida Sparse Matrix Collection)**, доступный по адресу [6], представляет собой обширную коллекцию разреженных матриц, которая используется для исследований в области численного анализа, оптимизации, и других областях науки и инженерии. Этот ресурс предлагает широкий набор данных, которые относятся к реальным примерам из различных приложений, включая проблемы структурного анализа, термического анализа, электромагнетизма и многих других.

2. **LDBC Social Network Benchmark (SNB)** [5] предоставляет наборы данных, которые моделируют социальные сети и предназначены для тестирования и оценки производительности графовых баз данных. Цель здесь – в создании стандартов и методологий для бенчмаркинга систем баз данных, что позволяет оценивать как скорость выполнения запросов к данным, так и другие аспекты производительности. Наборы данных LDBC включают в себя сценарии реального времени, такие как обработка динамически изменяющихся графов (например, в социальных сетях с добавлением/удалением друзей, сообщений и т.д.

3. **CFPQ_Data repository**, доступный по адресу [2], фокусируется на предоставлении данных для запросов на графах с ограничениями, задаваемыми формальными языками (Context-Free Path Querying - CFPQ). Это направление исследований активно развивается в области обработки запросов к графам и базам данных. Оно исследует возможности формирования запросов с ис-

пользованием контекстно-свободных грамматик для выявления сложных шаблонных связей между узлами графа, что оказывается полезным, например, в анализе биологических сетей, сетевой безопасности, исследованиях связности в социальных сетях и т.д.

В то время как SuiteSparse Matrix Collection фокусируется на предоставлении разреженных матриц для исследований в области численных методов и оптимизации, CFPQ_Data предоставляет данные, основанные на графах, для исследований в области запросов к графам и теории формальных языков.

Все ресурсы нацелены на исследователей и разработчиков, но SuiteSparse обращен, в первую очередь, к математикам, инженерам и специалистам в области численного моделирования, в то время как CFPQ_Data интересен исследователям в области теории формальных языков, компьютерных наук и смежных областей.

3. Реализация

Для начала, цель работы определена как создание системы хранения для результатов обработки различных графовых структур с последующей возможностью фильтрации и представления данных пользователям через интерфейс веб-сайта. В качестве основы для системы хранения были рассмотрены различные базы данных, включая PostgreSQL [7] и ClickHouse[1], с целью определения наиболее подходящего решения с точки зрения производительности, удобства использования и масштабируемо

Проведенное сравнительное исследование между PostgreSQL и ClickHouse позволило выявить ряд преимуществ PostgreSQL, в частности, его высокую надежность, поддержку транзакций и богатый функционал для работы с сложными запросами, что обеспечило выбор в пользу PostgreSQL.

Были созданы таблицы "graphs", "grammars" и "graphs_grammars" с целью эффективного моделирования сложных отношений между графиками и грамматиками. Связь многие-ко-многим была выбрана в качестве подхода к организации данных, поскольку позволяет устанавливать отношения между конкретными графиками и грамматиками.

Для каждой пары (граф, запрос), где граф может быть подвергнут запросу, необходимо хранить результаты в соответствующем виде. Например, для задачи достижимости нужно хранить множество вершин (список чисел). Таким образом, было предложено использовать таблицу для хранения ответов запросов, которая может быть подвергнута запросам фильтрации.

Например, предположим, что есть графы g1, g2, g3, и запросы q1, q2. Запрос q1 применим ко всем графам, в то время как запрос q2 применим только к g2 и g3. Для каждой комбинации (граф, запрос) необходимо хранить результаты, чтобы можно было проводить фильтрацию и выдавать результаты для всех отфильтрованных комбинаций. Примером таких комбинаций могут быть: (g1, q1), (g2, q1), (g3, q1), (g2, q2), (g3, q2).

Таким образом, связь многие-ко-многим используется для обеспечения эффективной организации данных и хранения результатов запро-

сов к графикам, а таблица для хранения результатов предоставляет возможность выполнять запросы фильтрации и получать необходимые результаты для различных комбинаций графиков и запросов. В дальнейшем, была настроена репликация данных [4] с использованием сервиса, аналогичного AWS Lambda, предоставляемого Yandex Cloud, для обеспечения высокой доступности и долговечности данных, а также гарантий их консистентности. А также использование сервиса Yandex S3 для хранения резервных копий и больших объектов, таких как архивы файлов. Это позволило эффективно организовать процесс репликации и синхронизации данных между различными хранилищами, обеспечивая высокую доступность и безопасность информации.

Таким образом, комбинация PostgreSQL, сервисов подобных AWS Lambda в Yandex Cloud и Yandex S3 предоставляет надежное и масштабируемое решение для управления данными, включая возможность хранения и работы с большими объектами, что делает его подходящим выбором для целей данного проекта.

Особое внимание в работе уделено разработке механизмов для эффективного хранения и обработки результатов запросов к графам. Как было упомянуто, при обработке графовых структур возникает необходимость в хранении результатов обработки в виде, поддерживающем быстрый доступ и эффективное использование данных. В рамках решения этой задачи был использован специализированный формат хранения, оптимизированный для хранения результатов, таких как множества достижимых вершин.

Дополнительно, в рамках данной работы была проведена модернизация веб-сайта, на котором результаты обработки графов представляются пользователям. Разработан и внедрен интерфейс для визуализации результатов в виде таблиц, что позволяет пользователям с легкостью фильтровать и анализировать интересующие их данные о графах и запросах. Интерфейс реализован с использованием современных технологий веб-разработки, что обеспечивает высокую скорость работы и удобство использования.

В результате проведенных исследований и разработок была создана

эффективная и удобная в использовании система для хранения, обработки и визуализации результатов обработки графовых структур, которая позволяет пользователям в реальном времени получать и анализировать данные о различных графах и выполненных над ними запросах. Таким образом, выполненная работа существенно улучшает возможности по работе с графовыми данными и предоставляет мощный инструмент для исследований в данной области.

Заключение

В ходе выполнения данной работы были выполнены следующие поставленные задачи:

- Выполнен обзор существующих подходов к хранению данных;
- Исследованы аналоги;
- Было создано хранилище данных с определенной архитектурой и ресурсами, спроектированными для определенного уровня нагрузки;
- Модификация сайта CPFQ_Data.

Продолжаются модификация процессов сбора, хранения и управления данными. Также ведется работа над апробацией алгоритмов анализа графов. Планируется добавление нового функционала в CFPQ_Data, а также добавление новых наборов графовых данных и модернизация архитектуры.

Список литературы

[1] ClickHouse Docs. — https://clickhouse.com/docs/ru. — Accessed: 2023-10-10.

[2] Formal Language Constrained Path Querying Data Repository. — https://formallanguageconstrainedpathquerying.github.io/CFPQ_Data/. — Accessed: 2024-04-05.

[3] Golfarelli M. Rizzi S. Data warehouse design: Modern principles and methodologies. — 2009. — ISBN: 0071610391.

[4] Kleppmann M. Designing Data-Intensive Applications. — 2017. — ISBN: 9781491903117.

[5] LDBC Dataset: SURF. — https://ldbcouncil.org/data-sets-surf-repository/. — Accessed: 2023-10-10.

[6] Matrix Market The SuiteSparse Matrix Collection (formerly the University of Florida Sparse Matrix Collection). — URL: http://sparse.tamu.edu/ (дата обращения: 16 сентября 2023 г.).

[7] PostgresPro. — URL: https://postgrespro.ru/docs (дата обращения: 16 января 2024 г.).

[8] R. Mattison. Data Warehousing and Data Mining for Telecommunications. — 1997.