Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.021:004.623

Ефимук Анастасия Андреевна

Алгоритм загрузки многомерных данных в денормализованное хранилище

АВТОРЕФЕРАТ

на соискание степени магистра технических наук
по специальности 1–40 80 02 «Системный анализ, управление и обработка информации»

_____

(подпись магистранта)

Научный руководитель
Гуринович Алевтина Борисовна
к.ф.-м.н., доцент

_____

(подпись руководителя)

Минск 2023

ВВЕДЕНИЕ

В современном мире компьютерные сети и вычислительные системы позволяют анализировать и обрабатывать большие массивы данных.

Ценность, надежность и достоверность знаний, полученных в результате интеллектуального анализа данных, зависит не только от эффективности используемых аналитических методов и алгоритмов, но и от того, насколько правильно подобраны и подготовлены исходные данные для анализа.

Большой объем информации усложняет поиск решений, но дает возможность получить намного более точный расчет и последующий анализ полученных решений. Поэтому, прежде чем приступать к анализу данных, необходимо выполнить ряд манипуляций с данными, цель которых — доведение данных до определенного уровня качества и информативности, а также организовать их интегрированное хранение в структурах, обеспечивающих их целостность, непротиворечивость, высокую скорость и гибкость выполнения аналитических запросов.

Одним из инструментов решения данных задач является процесс ETL — извлечение (Extract), преобразование (Transform) и загрузка (Load) данных. Данный процесс представляет собой комплекс операций, реализующих процесс переноса первичных данных из различных источников, таких как базы данных, файлы, API и т. д., в аналитическое приложение или поддерживающее его хранилище данных. Является составной частью этапа консолидации данных в обработке и анализе данных,

Данный процесс представляет собой комплекс операций, реализующих процесс переноса первичных данных из централизованного репозитория, используемого для приема и хранения больших объемов данных в их исходном виде (озера данных) в аналитическое приложение или поддерживающее его хранилище данных. Является составной частью этапа консолидации данных в и преобразования для адаптации к целевому формату

В ходе работы будут использоваться различные методы исследования:

– Методы моделирования и симуляции, чтобы создать виртуальную среду загрузки многомерных данных в денормализованное хранилище, чтобы оценить производительность и эффективность алгоритма

– Методика эмпирического анализа включает сбор и анализ реальных данных загрузки многомерных данных в денормализованное хранилище.

– Сравнительное исследование различных алгоритмов загрузки многомерных данных в денормализованное хранилище. Это включает сравнение производительности, эффективности и точности разных алгоритмов в разных сценариях загрузки и на разных типах данных.

Научная новизна разработки состоит в том, чтобы исследовать новые аспекты и проблемы, связанные с загрузкой многомерных данных в денормализованное хранилище. Предложены новые решения, которые заключаются в ускорении обработки данных, а также реализован инструмент, который ускоряет загрузку данных в денормализованное хранилище.

Цель исследования. Целью магистерской диссертации является изучение методов и алгоритмов работы с данными и разработка алгоритма оптимального подхода к загрузке и преобразованию данных для последующего использования в аналитических целях.

Задачи исследования:

– исследование различных алгоритмов и методов загрузки многомерных данных в денормализованное хранилище и сравнение их производительности, эффективности и точности;

– исследование методов и техник оптимизации загрузки многомерных данных, чтобы снизить время загрузки, увеличить пропускную способность и справиться с большими объемами данных;

– исследование методов управления ошибками при загрузке многомерных данных и возможности отката в случае неудачной загрузки;

– исследование методов и техник загрузки данных из различных источников, таких как базы данных, файлы CSV или Excel, API и другие источники;

– исследование методов загрузки многомерных данных в денормализованное хранилище при работе с большими объемами данных и распределенными системами;

– исследование методов оценки качества данных при загрузке в денормализованное хранилище.

Объект исследования. Объектом исследования являются алгоритмы и методологии для обработки и загрузки многомерных данных.

Предмет исследования. Предметом исследования является алгоритм, поддерживающий эффективную загрузку многомерных данных в хранилище.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель исследования

Целью диссертации является определение оптимального подхода к загрузке и преобразованию данных для последующего использования.

Задачи исследования:

– обзор систем и алгоритмов преобразования;

– анализ существующих методов и алгоритмов трансформации данных;

– модернизация алгоритма, на его основе.

Объектом исследования являются алгоритмы трансформации и последующая загрузка данных в денормализованное хранилище.

Личный вклад магистранта

Магистрантом выполнена работа в полном объеме. Постановка задач и обсуждение результатов проводились совместно с научным руководителем и сотрудниками кафедры информационных технологий автоматизированных систем Белорусского государственного университета информатики и радиоэлектроники. Соавторы опубликованных работ принимали участие в обсуждении промежуточных и конечных результатов. Обработка, интерпретация данных, а также выводы сделаны автором самостоятельно.

Апробация результатов диссертации

Основные положения диссертационной работы докладывались на следующих научных конференциях (по результатам исследования опубликовано 9 работ):

– Proceedings of the VI International Scientific and Practical Conference, Warsaw, September 30, 2018;

– Proceedings of the X International Scientific and Practical Conference, February 28, 2018. - Warsaw: RS Global Sp. z O.O;

– International Academy Journal Web of Scholar. – 2019. Warsaw: RS Global Sp. z O.O;

– The International Conference on Information Technologies and Systems ITS 2020/2021 (Минск 2021);

– XIII Международная научно-практическая конференция профессорско-преподавательского состава, аспирантов, магистрантов и студентов, БИП — Университет права и социально-информационных технологий, 2023;

– XXVIII Республиканский конкурс научных работ студентов, Минск 2022 (диплом 1-ой степени);

– 56-я научная конференция аспирантов, магистрантов и студентов БГУИР;

– 58-я научная конференция аспирантов, магистрантов и студентов БГУИР;

– 59-я научная конференция аспирантов, магистрантов и студентов БГУИР.

Структура и объем диссертации

Диссертация состоит из оглавления, общей характеристики работы, введения, четырёх глав, заключения, списка использованных источников. Полный объём диссертации составляет 59 страниц, включая 12 рисунков и одно приложение. Список использованных источников включает 25 наименований и занимает 2 страницы.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Введение предоставляет общую информацию о контексте исследования, обоснование его актуальности, описание методов исследования и обозначение научной новизны работы. Оно также формулирует цель и задачи исследования, а также определяет объект и предмет исследования.

В **первой главе** рассматривается актуальность исследования в области хранилищ данных, а также преимущества использования хранилищ данных. Описываются различные аспекты создания и управления хранилищем данных, включая интеграцию данных, обработку больших объемов. Также рассматриваются основные задачи исследования.

Во **второй главе** производится рассмотрение методов работы с многомерными данными. В главе приводится детальное описание технологий загрузки и обработки данных. Предоставляется информация о методах трансформации при работе с многомерными данными.

В **третьей главе** описан процесс модификации первоначальной версии алгоритма загрузки многомерных данных. Проводится обзор инструментов для реализации модифицированной версии описываемого алгоритма. Кроме того, предоставляется информация об анализе производительности новой версии алгоритма.

В **четвёртой главе** описывается процесс апробации алгоритма, детально описывается архитектура разработанной системы. Кроме того, подробно отражена информация о поведении алгоритма в нештатных ситуациях.

В **пятой главе** рассматривается один из практических примеров использования алгоритма, предоставлены результаты работы алгоритма, описываются основные выводы по результатам исследования. В главе подробно описываются этапы решения задачи и на основании результатов эксперимента делается вывод о целесообразности использования данного алгоритма.

ЗАКЛЮЧЕНИЕ

В диссертационном исследовании был предложен алгоритм загрузки многомерных данных в денормализованное хранилище. Рассмотрены основные принципы работы, приведены примеры использования и сравнения с аналогами. Тестирование алгоритма показали его эффективность.

Были проанализированы требования к алгоритму, проведено исследование существующих аналогов и разработана улучшенная версия алгоритма. Для его реализации была использована комбинация инструментов, включая Python, PostgreSQL, Argo и другие.

Проведенное тестирование показало, что разработанный алгоритм обладает высокой производительностью, а также улучшенным качеством данных и повышенной надежностью.

Кроме того, были предложены возможные пути улучшения алгоритма, которые могут быть реализованы в будущих исследованиях.

Одним из инструментов решения данных задач является процесс ETL. Данный процесс представляет собой комплекс операций, реализующих процесс переноса первичных данных из различных источников в аналитическое приложение или поддерживающее его хранилище данных. Является составной частью этапа консолидации данных в анализе данных.

ETL-системы играют жизненно важную роль в системе загрузки и агрегации данных. Они обеспечивают аналитикам и ученым доступ к данным из нескольких систем приложений.

Однако, для максимальной эффективности алгоритма, необходимо учитывать особенности конкретного проекта и проводить тщательное тестирование перед внедрением.

Изучив все аспекты работы алгоритма, можно сделать вывод о его эффективности и удобстве использования. Алгоритм показал себя как надежный инструмент для загрузки многомерных данных, способный обеспечивать высокую скорость работы и качество результата.

Таким образом, можно сделать вывод о том, что алгоритм загрузки многомерных данных является эффективным и удобным инструментом для работы с данными, который может быть применен в различных проектах и обеспечивать высокий уровень качества результата.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

[1-A] Молодежный взгляд на подходы и тенденции в области дистанционного обучения / Зубович В.О., Христофорова А.А.// International Trends in Science and Technology: Proceedings of the VI International Scientific and Practical Conference, Warsaw, September 30, 2018. – Warsaw, 2018. – Vol.1. – P. 8 – 11.

[2-A] Программные продукты, предназначенные для решения задач современной экономики на платформе 1С: Предприятие / Бакунова О. М, Христофорова А.А. // International Trends in Science and Technology: Proceedings of the X International Scientific and Practical Conference, February 28, 2018. – Warsaw: RS Global Sp. z O.O. – Vol. 1. – P. 7 – 13.

[3-A] Автоматизация задач банковского обслуживания / Бакунова О. М., Христофорова А. А. // International Academy Journal Web of Scholar. – 2019. – № 10 (40). – С. 36-38. – DOI: 10.31435/rsglobal_wos/31102019/6739.

[4-A] Христофорова, А. А. Web-приложение администратора компании «Taqtile» на языке Angular JS / Христофорова А. А. // Информационные системы и технологии: материалы 56-й научной конференции аспирантов, магистрантов и студентов, Минск, 18–20 мая 2020 г. / Белорусский университет информатики и радиоэлектроники, Институт информационных технологий; редкол.: А. А. Охрименко. – Минск: БГУИР, 2020. – С. 78–80.

[5-A] Клюцук, А. С. Роль Республиканского центра обработки данных в жизни общества и в развитии облачных технологий / А. С. Клюцук, А. А. Христофорова // Электронные системы и технологии : сборник тезисов докладов 56-ой научной конференции аспирантов, магистрантов и студентов БГУИР, Минск, 18 – 20 мая 2020 г. / Белорусский государственный университет информатики и радиоэлектроники. – Минск, 2020. – С. 119.

[6-A] Христофорова, А. А. Алгоритм загрузки многомерных данных в денормализованное хранилище / Христофорова А. А., Гуринович А. Б. // Информационные технологии и системы 2021 (ИТС 2021) = Information Technologies and Systems 2021 (ITS 2021) : материалы международной научной конференции, Минск, 24 ноября 2021 г. / Белорусский государственный университет информатики и радиоэлектроники ; редкол.: Л. Ю. Шилин [и др.]. – Минск, 2021. – С. 228–229.

[7-А] Христофорова А. А. Клиент-серверное приложение администратора платформы Manifest / Христофорова А.А. // XXVIII Республиканский конкурс научных работ студентов, Минск, 2021 г.

[8-А] Христофорова, А. А. Алгоритм загрузки многомерных данных в денормализованное хранилище / А. А. Христофорова // Информационные технологии и управление : материалы 58-ой научной конференции аспирантов, магистрантов и студентов, Минск, 18–22 апреля 2022 года / Белорусский государственный университет информатики и радиоэлектроники ; редкол.: Л. Ю. Шилин [и др.]. – Минск, 2022. – С. 55–56.

[9-А] Никульшина К.Б., Христофорова, А. А. Использование нейронных сетей при загрузке многомерных данных в денормализованное хранилище/ К.Б.Никульшина, А. А. Христофорова // Информационные технологии и управление : материалы 59-ой научной конференции аспирантов, магистрантов и студентов, Минск, 17–21 апреля 2023 года / Белорусский государственный университет информатики и радиоэлектроники ; редкол.: М. Л. Маковский [и др.]. – Минск, 2023.