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

УДК 004.65

ЖУКОВ
Алексей Максимович

АЛГОРИТМЫ ОБРАБОТКИ БОЛЬШИХ ОБЪЕМОВ ИНФОРМАЦИИ В ХРАНИЛИЩАХ ДАННЫХ

АВТОРЕФЕРАТ
на соискание степени магистра технических наук
по специальности 7-06-0612-03— Системы управления информацией

Научный руководитель
к.т.н., доц.
Навроцкий Анатолий
Александрович

Минск, 2025

ВВЕДЕНИЕ

Современные корпоративные и аналитические информационные системы характеризуются непрерывным ростом объёмов данных и повышенными требованиями к их актуальности. В условиях цифровизации бизнес-процессов и широкого распространения транзакционных систем особую роль приобретает задача оперативной и надёжной доставки изменений данных в аналитические хранилища. Одним из ключевых механизмов, обеспечивающих решение данной задачи, являются CDC-конвейеры (Change Data Capture), позволяющие в потоковом режиме фиксировать и применять изменения, происходящие в источниках данных.

Важнейшей составляющей CDC-driven ETL является стратегия записи данных в аналитическое хранилище. От выбранной стратегии напрямую зависят задержка обновления данных, пропускная способность системы, устойчивость обработки и эффективность использования вычислительных ресурсов. В современных системах обработки данных широко применяются методы микробатчинга, различные режимы записи и алгоритмы компрессии, каждый из которых обладает своими преимуществами и ограничениями. Однако в условиях динамически изменяющейся нагрузки статический выбор параметров обработки зачастую приводит либо к росту задержки, либо к деградации и отказам системы.

Разработка новых и совершенствование существующих алгоритмов обработки потоков изменений данных представляет собой значимый научный и практический интерес. Оптимизация стратегий записи и параметров обработки позволяет не только повысить производительность аналитических хранилищ, но и способствует развитию распределённых вычислительных моделей, потоковых платформ и архитектур больших данных. Результаты подобных исследований находят применение в корпоративных аналитических системах, финансовых и торговых платформах, системах мониторинга и отчётности, где требования к актуальности данных являются критическими.

Актуальность темы диссертационного исследования определяется необходимостью создания адаптивных и устойчивых алгоритмов обработки CDC-потоков, способных обеспечивать минимальную задержку обновления данных и стабильную работу системы при существенно меняющихся профилях нагрузки, в том числе при burst-нагрузках и ограниченных ресурсах вычислительной инфраструктуры.

Целью диссертационного исследования является разработка адаптивного алгоритма оптимизации стратегии записи данных в аналитическое хранилище в рамках CDC-driven ETL-конвейера,

обеспечивающего снижение end-to-end задержки обработки и повышение устойчивости системы при динамически изменяющейся нагрузке.

Для достижения поставленной цели в диссертационной работе были поставлены и решены следующие задачи:

— выполнен анализ алгоритмов обработки больших объёмов данных и существующих подходов к построению CDC-driven ETL-конвейеров;

— исследованы преимущества и ограничения статических стратегий записи, микробатчинга и алгоритмов компрессии в аналитических хранилищах;

— разработан адаптивный алгоритм выбора стратегии записи и параметров обработки данных на основе текущей интенсивности входного потока;

— реализован экспериментальный стенд для оценки эффективности предложенного алгоритма и выполнено сравнение с базовыми режимами обработки;

— проведена экспериментальная оценка влияния адаптивного подхода на задержку обработки, устойчивость выполнения и поведение системы при различных профилях нагрузки.

Достоверность полученных научных результатов подтверждается согласованностью теоретических положений с результатами вычислительных экспериментов, воспроизводимостью полученных данных, а также практической реализацией предложенного алгоритма в рамках потоковой платформы обработки данных и аналитического хранилища.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования.

Целью магистерской диссертации является разработка алгоритма адаптивной обработки потоков изменений данных в аналитическом хранилище, обеспечивающего снижение задержки обработки и повышение устойчивости системы при работе с большими объёмами информации.

Для достижения поставленной цели в работе необходимо решить следующие задачи:

— выполнить анализ существующих подходов к обработке больших объёмов данных в аналитических хранилищах, в том числе в рамках CDC-driven ETL-конвейеров;

— исследовать преимущества и недостатки статических стратегий обработки данных с точки зрения задержки, пропускной способности и устойчивости системы;

— разработать алгоритм адаптивного выбора стратегии записи данных в аналитическое хранилище в зависимости от текущей нагрузки;

— реализовать разработанный алгоритм и провести экспериментальное исследование его эффективности на различных профилях нагрузки, включая пиковые режимы.

Объектом исследования в данной магистерской диссертации являются системы обработки больших объёмов данных и аналитические хранилища.

Предметом исследования является алгоритм адаптивной обработки потоков изменений данных и выбора стратегии записи в аналитическом хранилище.

Личный вклад соискателя.

Все результаты и положения, выносимые на защиту, получены автором лично. Научный руководитель принимал участие в постановке задачи, определений возможных путей из решений, в предварительном анализе, обсуждении результатов и практических исследований, проведенных автором лично.

Научная новизна.

Научная новизна работы заключается в разработке алгоритма адаптивного управления параметрами обработки потоков изменений данных, который сочетает в себе динамический выбор стратегии записи, параметров микробатчинга и алгоритмов компрессии. В отличие от существующих решений, использующих статические конфигурации, предложенный алгоритм учитывает текущую интенсивность входного потока и эмпирически

определённые пороговые значения, что позволяет минимизировать задержку обработки и предотвратить деградацию системы при высоких нагрузках. Полученные в ходе экспериментального исследования результаты подтверждают эффективность предложенного подхода.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Диссертационная работа состоит из введения, общей характеристики работы, трех глав основной части, заключения и библиографического списка, оформленного в порядке упоминания источников. Общий объем диссертации составляет 63 страницы, из них 2 страницы занимает библиографический список.

Во введении обоснована актуальность темы исследования, определены цель и задачи работы, объект и предмет исследования, а также раскрыты научная новизна и практическая значимость полученных результатов. Кратко охарактеризованы основные проблемы обработки больших объемов данных в аналитических хранилищах, связанные с ростом нагрузки, изменчивостью профилей поступления данных и ограничениями статических конфигураций ETL-конвейеров.

Первая глава «Обзор существующей литературы» посвящена анализу научных и практических работ, посвящённых обработке больших объёмов данных в аналитических хранилищах. В разделе рассмотрены алгоритмы обработки больших объёмов данных в хранилищах данных, а также основные архитектуры хранилищ и модели обработки, применяемые в современных системах. Отдельное внимание уделено распределённым вычислительным моделям обработки данных, их особенностям и ограничениям при работе с потоками изменений. Проанализированы подходы к оптимизации обработки данных в современных хранилищах, включая методы повышения пропускной способности и снижения задержки обработки. В заключительной части главы рассмотрены методы инкрементальной и адаптивной обработки больших объёмов данных, что позволило выявить существующие ограничения статических конфигураций и обосновать необходимость разработки адаптивных алгоритмов управления параметрами обработки данных.

Вторая глава «Теоретические основы адаптивной инкрементальной обработки данных в хранилищах» содержит формализацию задачи оптимизации параметров обработки данных в CDC-driven ETL-конвейерах. В главе рассмотрены модели формирования задержки обработки, компромисс между задержкой и пропускной способностью, а также влияние параметров микробатчинга, алгоритмов компрессии и стратегий применения изменений на характеристики системы. Введены управляющие параметры обработки и обоснована необходимость перехода от статических конфигураций к адаптивным алгоритмам управления, способным учитывать текущие характеристики входного потока данных и состояние вычислительных ресурсов.

Третья глава «Экспериментальное исследование и реализация адаптивного алгоритма обработки данных» посвящена практической реализации и экспериментальной проверке предложенного алгоритма. В главе описана архитектура экспериментального стенда на основе потоковой платформы Apache Flink и аналитического хранилища Apache Iceberg, приведена методика измерения end-to-end задержки и используемые статистические показатели эффективности. Представлены результаты экспериментальных исследований при различных профилях нагрузки, включая малые, средние и высокие объемы данных, а также проведен сравнительный анализ статических и адаптивных режимов обработки. Показано, что предложенный алгоритм позволяет снизить задержку обработки и повысить устойчивость системы при пиковых нагрузках.

В заключении сформулированы основные результаты диссертационной работы, приведены выводы по достигнутым целям и задачам исследования, а также обозначены направления дальнейшего развития предложенного алгоритма и возможные области его практического применения.

ЗАКЛЮЧЕНИЕ

В рамках данной магистерской диссертации была рассмотрена задача повышения эффективности обработки больших объёмов данных в аналитических хранилищах в условиях динамически изменяющейся нагрузки. Актуальность выбранной темы обусловлена широким распространением CDC-driven ETL-конвейеров и возрастающими требованиями к минимизации задержки обработки данных при одновременном обеспечении устойчивости и масштабируемости систем.

Проведённый анализ современных подходов к организации потоковой обработки данных и CDC-driven ETL показал, что в большинстве практических реализаций параметры микробатчинга, компрессии и стратегии применения изменений задаются статически, что существенно ограничивает эффективность системы при изменении профиля нагрузки.

На основе сформулированной модели обработки CDC-потока выявлены фундаментальные ограничения статических конфигураций и обоснована необходимость перехода к адаптивным методам управления параметрами обработки данных. Введён вектор управляющих параметров, включающий размер микробатча, стратегию записи и алгоритм компрессии, а также показано, что оптимальные значения этих параметров зависят от текущего состояния системы и характеристик входного потока.

Была разработана и экспериментально исследована реализация CDC-конвейера на базе Kafka/Debezium, Apache Flink и Apache Iceberg. Предложена методика измерения end-to-end задержки обработки, включающая использование статистических показателей средней и хвостовой латентности. Проведена серия экспериментов при различных профилях нагрузки — от малых серий событий до burst-нагрузок объёмом сотни тысяч записей. Установлено, что статические стратегии записи обладают ограниченной областью применимости: режим LL(Low Latency) обеспечивает минимальную задержку при малой нагрузке, но теряет устойчивость при высоких объёмах данных, тогда как режим HT (High Throughput) демонстрирует стабильность, но сопровождается увеличением задержки обработки.

Предложена реализация адаптивного алгоритма выбора стратегии записи и параметров обработки данных, основанная на динамической оценке интенсивности входного потока. Алгоритм реализован в виде управляющего слоя с использованием служебной таблицы и механизма гистерезиса, что позволило избежать дрожания режимов и сохранить exactly-once семантику обработки. Экспериментальная эмуляция адаптивного режима показала, что предложенный алгоритм обеспечивает более низкую среднюю и хвостовую задержку по сравнению со статическим режимом HT, сохраняя при этом

устойчивость системы при высоких нагрузках, на которых режим LL становится неработоспособным.

Разработанный алгоритм адаптивного управления параметрами CDC-driven ETL-конвейера, позволил повысить эффективность обработки больших объёмов данных в аналитическом хранилище, что позволяет его эффективно использовать при проектировании и эксплуатации ETL-систем с динамически изменяющейся нагрузкой.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ

1. А. М. Жуков, А. А. Навроцкий // Информационные технологии и системы 2025 (ИТС 2025): материалы Международной научной конференции, Минск, БГУИР, 2025 - С. 199–200.