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

УДК 004.62

Асташевич
Роман Григорьевич

Методика и программно-алгоритмическое обеспечение динамического формирования отчетов на основе данных OLAP куба

АВТОРЕФЕРАТ
на соискание степени магистра технических наук

по специальности 1-40 80 04 Математическое моделирование, численные методы и комплексы программ

_____

Научный руководитель
В.Г. Лукьянец, кандидат
технических наук, доцент

_____

Минск 2017

КРАТКОЕ ВВЕДЕНИЕ

В наше время без систем управления базами данных не обходится практически ни одна организация, все они собирают и хранят в своих базах гигабайты данных о клиентах, продуктах и сервисах. Такие базы данных называют операционными или транзакционными, поскольку они характеризуются огромным количеством небольших транзакций, или операций записи-чтения. По мере сбора транзакций суммарные значения меняются очень быстро, поэтому два анализа, проведенные с интервалом в несколько минут, могут дать разные результаты. Кроме того, необходимые для анализа данные могут храниться в нескольких системах.

Некоторые виды анализа требуют таких структурных изменений, которые недопустимы в текущей оперативной среде. Этим объясняется интерес к объединению и анализу данных учетной системы с помощью технологии OLAP (On-Line AnalyticalProcessing). OLAP – технология обработки данных, заключающаяся в подготовке суммарной (агрегированной) информации на основе больших массивов данных, структурированных по многомерному принципу.

Основными трудностями при работе с многомерными массивами данных, как правило, являются их разреженность и требование к долговременному хранению набора. Долгое время задача организации хранения и управления данными решался с использованием реляционных баз данных. Этот подход прекрасно подходит для данных, представимых в виде табличной структуры, позволяя при этом выполнять запросы с использованием традиционного механизма работы с реляционными СУБД – языка SQL.

Порой возникают ситуации, где мощь данной стратегии является избыточной модель данных не совсем ложится в привычное табличное представление. В этом случае предпочтительнее использование альтернативных специализированных технологий, которые не хранят данные в

виде таблиц, а предоставляют возможность оперировать их представлением в произвольном формате.

В диссертациирассмотрены вопросы проектирования и реализации системы хранения типа «ключ-значение», являющейся по своей сути подмодулем хранилища аналитических данных и системой кэширования информации в памяти с возможностью выполнения запросов для получения агрегированных данных в составе системы динамического формирования отчетов на основе данных OLAP куба.

Задачи, которые ставятся перед модулем следующие:

а) проверка наличия данных, соответствующих определенному ключу;

б) обработка запросов подсчета агрегатов;

в) обработка запросов получения метаданных хранимых записей;

г) возможность реализации в рамках модуля препроцессинга и трансформации данных по предопределенным правилам перед вставкой в кэш-хранилище;

д) поддержка операции добавления данных (append) без необходимости перегенерации;

е) выборочное обновление хранимых данных.

Основными недостатками существующих систем хранения данных типа «ключ-значение» являются: медленная вставка при большом количестве хранимых записей, плохая работа в конкурентной среде, недостаточные гарантии персистентности, отсутствие возможности работы в кластере и избыточная сложность при решении типовых задач, поставленных перед данным модулем.

Данная тема является актуальной, поскольку задача подготовки суммарной (агрегированной) информации на основе больших массивов данных, структурированных по многомерному принципу, лежит в основе многих корпоративных систем построения отчетов.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Целью данной работы ставится анализ представленных на рынке решений, рассмотрение основных моментов при проектировании и разработке консистентного, надежного хранилища данных.

Задачами исследования являются:

- анализ существующих хранилищ данных;
- определение слабых и сильных сторон существующих решений;
- проектирование хранилища «ключ-значение» с учетом проведенного исследования;
- разработка интерфейсов взаимодействия с другими модулями;

Объект исследования – программное и аппаратное обеспечение системы хранения данных в памяти.

Предмет исследования – архитектура сервисов хранения данных, размещающих информацию по типу ассоциативного словаря в оперативной памяти.

Результаты, приведенные в диссертации, получены соискателем лично. Вклад научного руководителя В.Г. Лукьянеца, заключается в формулировке целей и задач исследования.

Диссертация состоит из введения, трех глав, заключения, библиографического списка и списка публикаций автора В первой главе представлен анализ предметной области, приведена классификация OLAP-систем а также дана сравнительная характеристика программных реализаций, сформулирована общая постановка на исследование. Вторая глава посвящена разработке архитектуры системы, и обзору используемых технологий. В третьей главе предложена практическая реализация программного средства.

Общий объем работы составляет 69 страниц, из которых основной текст – 59 страниц, в том числе21рисунка на 18 страницах, 8 таблиц на 9 страницах и список использованных источников из 31 наименования на 4 страницах.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность диссертационной работы, приведены цель и основные задачи работы.

В первой главе «Анализ моделей и программных средств анализа данныхв OLAP технологии» приводятся основные понятия предметной области, приведена концепция OLAP-систем и дана их классификация, приведен обзор основных программных реализаций.

Во второй главе приводится обзор архитектурных решений используемых в OLAP-системах. Осуществляется обзор технологий, позволяющих решить поставленную задачу, по разработке хранилища данных типа «ключ-значение», максимально эффективно. Рассматриваются механизмы организации хранения данных в различных регионах памяти предоставляемые JDK. Производится тест производительности различных способов хранения данных. Тест заключается в следующем: производится вставка 50 миллионов простейших кортежей в хранилище, а затем происходит подсчет агрегата по вставленным ранее векторам. Результаты сравнений представлены на рисунке 1. Тесты проводились на MacBookPro и результаты могут отличаться от запуска на других машинах.

Benchmark	Mode	Samples	Mean	Mean error	Units
StorageBenchmark.linkedList	avgt	5	21554.219	26206.010	ms/op
StorageBenchmark.arrayList	avgt	5	4309.324	5777.266	ms/op
StorageBenchmark.byteBuffer	avgt	5	1778.355	220.249	ms/op
StorageBenchmark.mappedByteBuffer	avgt	5	1132.380	1201.578	ms/op
StorageBenchmark.directByteBuffer	avgt	5	766.686	215.527	ms/op
StorageBenchmark.unsafe	avgt	5	483.926	26.336	ms/op
StorageBenchmark.noop	avgt	5	17.313	0.650	ms/op

Рисунок 2.1 – Сравнение различных методов реализации простейшего хранилища

Производится сравнение различных механизмов сериализации и транспортировки дынных. Ниже в таблице 1 представлен результат тестов 1000 обращений клиент-сервер, расположенных на разных машинах.

Таблица 1 – Сравнение Thrift и GRPC реализаций

Формат сериализации	avg(1 мин,сбщ./сек)	avg(5 мин, сбщ./сек)	avg(15 мин, сообщ./сек)	avg(сбщ./сек)
Thrift+TCompactProtocol	1.12	1.13	1.07	1.11
GRPC+Protobuf3	1.68	1.65	1.63	1.67

Для организации ETL-потока используется конструкция на основе Disruptor.Для обеспечения быстрого и прозрачного механизма трансформаций, который бы эффективно масштабировался на многопроцессорных системах, при помощи направленного графа обработки был реализован многоэтапный механизм (рисунок 2).

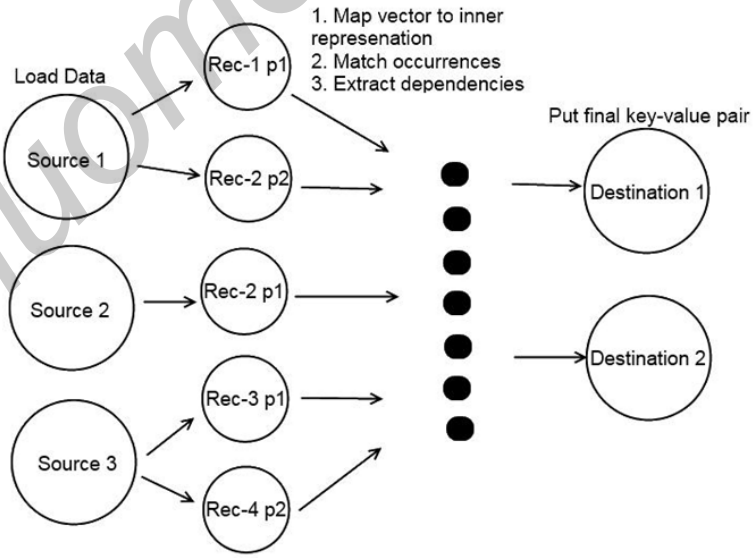

Рисунок 2 – Примерный порядок организации ETL

В третьей главе «Разработка структуры программно-алгоритмического обеспечения хранилища данных для динамического формирования отчетов»

описана архитектура разработанного хранилища типа «ключ-значение». Описаны разработанные модели хранения данных и базовых примитивов хранилища данных – индексов.

Для координации доступа ко всем индексам в системе существует фасад – IndexManager. Данный класс позволяет абстрагироваться и упростить пользователю API кэша вставку элементов и поиск по заданным фильтрам. Асимптотика операций на реализованных индексах представлена в таблице 3.

Таблица 3 – Асимптотика операций на индексах

Операция	Лучший	Средний	Худший
Вставка	O(N)	O(N)	O(N)
Поиск	O(logN)	O(logN)	O(logN)
Удаление	O(N)	O(N)	O(N)
Объединение	O(N^3)	O(N^3)	O(N^3)

Для оптимизации проверки наличия определенного значения в индексах использовалась вероятностная структура данных фильтр Блума (рисунок 3).

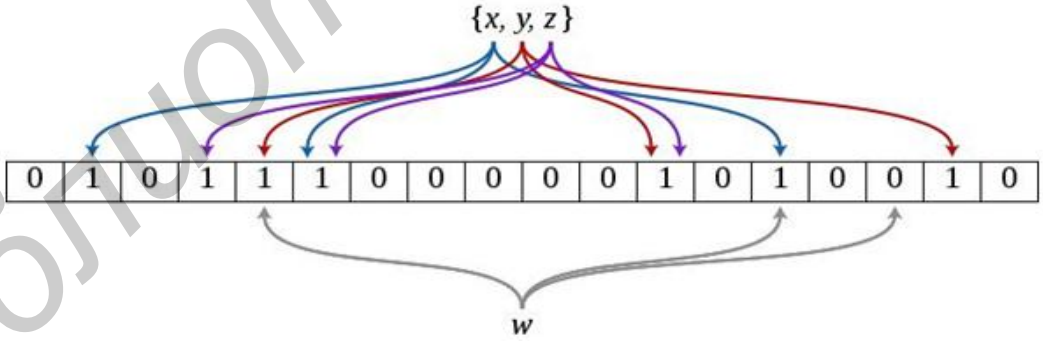

Рисунок 3.11 – Фильтр Блума

Использование фильтра Блума позволило увеличить производительность операции поиска на индексах, в которых отсутствовали искомые ключи.

ЗАКЛЮЧЕНИЕ

Данная работа посвящена актуальному направлению проектирования систем хранения данных, а именно систем типа «ключ-значение». Упомянутый класс хранилищ важен при эффективном решении задачах бизнес-аналитики и построении DWH-систем. Предоставляя функциональность кэширования и обеспечивая высокопроизводительный доступ к данным, разработанное хранилище позволяет снизить затраты, возникающие при неэффективной реализации систем построения отчетности.

В процессе работы были получены следующие результаты:

а) проведен анализ существующих хранилищ данных;

б) определены сильные и слабые стороны существующих решений;

в) разработаны компоненты для системы построения отчетов на основе проведенного исследования;

г) разработаны интерфейсы взаимодействия с другими компонентами системы.

Результатом проделанной работы является созданное и функционирующее в реальных условиях хранилище типа «ключ-значение», эффективно использующее вычислительные ресурсы и предоставляющее возможность качественно решать поставленные задачи.

Являясь компонентом системы построения отчетности, продукт, основанный на исследованиях, изложенных в данной работе, существенно изменил качество обслуживания клиентов по сравнению с предыдущим решением на основе реляционных СУБД. Можно утверждать, что полученные результаты исследований нашли реальное применение и были внедрены на практике в виде работающего модуля. Задачи, перечисленные во введении к данной работе, были успешно выполнены, а поставленная цель достигнута.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1-А. Асташевич Р.Г. РАЗРАБОТКА DWH-СИСТЕМЫ ТИПА «КЛЮЧ-ЗНАЧЕНИЕ»//Проблемы современных интеграционных процессов и пути их решения: сборник статей международной научно-практической конференции: в двух частях, Россия, г. Омск, 13 декабря 2016 г. – Томчасть 2, С. 45 – 48