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

УДК 004.021:004.67

Гобрик
Олег Дмитриевич

АЛГОРИТМЫ ОБРАБОТКИ ДАННЫХ В ПАРАЛЛЕЛЬНЫХ СИСТЕМАХ

АВТОРЕФЕРАТ
на соискание степени магистра технических наук

по специальности 1–40 80 02 «Системный анализ, управление и обработка информации»

Научный руководитель
Гуринович Алевтина Борисовна
кандидат физ.-мат. наук, доцент

Минск 2021

ВВЕДЕНИЕ

Рост объема и источников данных требуют изменения подходов и технологий обработки. Реляционные СУБД остаются основной технологией управления данными для обработки структурированных данных. Реляционные базы данных хранят сущности в отдельных таблицах, которые обычно хорошо нормализованы. Эта структура удобна для операционных баз данных, но сложные многотабличные запросы в ней выполняются относительно медленно.

В настоящее время базы данных и приложения по оперативной аналитической обработке (OLAP) — основное звено для поддержки принятия решений. Как правило, серверы OLAP реализованы в верхней части собственной системы на основе массивов хранения или как расширение для традиционных реляционных СУБД.

Для оптимизации обработки используются технологии MapReduce и NoSQL. На данном этапе развития технологии нет математических методов оценки времени выполнения запросов к базам данных на платформе распределенных параллельных вычислений на этапе проектирования системы. Методы должны учитывать специфику сложных запросов к базе данных, используемых в процессе построения аналитических систем.

В исследовании решаются следующие задачи:

– выполнение анализа предметной области;

– исследование новейших методик и трендов в области обработки данных;

– анализ и сравнение методов доступа к данным в параллельных распределённых базах данных на платформе MapReduce/Spark;

– разработка метода доступа к данным на базе MapReduce/Spark;

– модифицировать методы работы со сложными и многомерными данными и продемонстрировать ее эффективность на примере;

В первой главе проводится анализ предметной области и основы параллельных систем.

Во второй главе рассматриваются основные подходы, модели и технологии для параллельной обработки данных.

В третьей главе описываются модифицированные методы обработки данных. Предложен алгоритм для автоматической генерации программы выполнения запроса в параллельной распределённой среде и анализируется его эффективность на тестовой базе данных.

Магистерская диссертация выполнена самостоятельно, проверена в системе «Антиплагиат». Процент оригинальности соответствует норме, установленной кафедрой.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальности исследования

Распределённые информационные системы, построенные на базе вычислительных и телекоммуникационных сетей, занимают в современном мире все более значимое место за счет расширения сферы применения, повышения качества обслуживания пользователей. Потребность в таких системах постоянно возрастает, как и требования, предъявляемые к их характеристикам.

Из-за увеличения размерности информационных систем, происходит усложнение приложений, обрабатывающих запросы пользователей, решение проблемы управления нагрузкой распределенных устройств – актуальная задача, имеющая прикладное значение.

Цель исследования

Цель диссертационной работы модификация методов и алгоритмов для организации параллельной обработки данных на основе исследования и анализа существующих технологий.

Задачи исследования

1. Обзор существующих систем параллельной обработки данных;
2. Анализ существующих методов и алгоритмов параллельной обработки данных;
3. Реализация и исследование программного средства разработанного алгоритма.

Новизна полученных результатов

Научная новизна заключается в том, что был предложен метод представления SQL-запроса в виде вложенных деревьев. Разработан метод выполнения запроса с использованием фильтра Блума, основанном на последовательной реализации деревьев и использования фильтров Блума.

Личный вклад соискателя

Соискателем работа велась полностью в соответствии с индивидуальным планом магистерской диссертации. Все промежуточные результаты обсуждались на заседаниях кафедры информационных технологий автоматизированных систем Белорусского государственного университета информатики и радиоэлектроники. Результаты работы опубликованы в трёх конференциях.

Апробация результатов диссертации

Основные положения диссертационной работы докладывались на трёх научных конференциях:

– The International Conference on Information Technologies and Systems ITS 2019 (Минск 2019);

– Информационные технологии и управление: научная конференции аспирантов, магистрантов и студентов (Минск 2020 и 2021).

Основные положения, выносимые на защиту

1. Показано на основе анализа предметной области, что существующие методы нуждаются в модификации.

2. Предложена модификация метода с использованием фильтра Блума для выполнения запросов к распределенному хранилищу данных в среде MapReduce/Spark для повышения производительности выполнения анализа данных.

СОДЕРЖАНИЕ РАБОТЫ

Диссертация состоит из введения, общей характеристики работы, трех глав, заключения, списка использованных источников, приложения и листа графического материала.

В первой главе «анализ предметной области» дан краткий обзор развития параллельных систем. Описана последовательная работа процессов для обеспечения функционирования таких систем. Разобраны структуры параллельных систем. Рассмотрен параллелизм на уровне команд и на уровне процессоров.

Во второй главе рассматривается технология параллельной распределенной обработки данных MapReduce, основные составляющие: Map, Shuffle и Reduce. В главе приводятся преимущества и недостатки этой технологии. Проведен критический анализ существующих методов реализации соединения таблиц измерений и фактов, выявлены их общие недостатки. Выполнен анализ математических методов моделирования процессов доступа к базам данных, перечисляются их недостатки.

В третей главе рассматриваются существующие схемы баз данных. Показаны способы представления SQL-запроса в виде вложенных деревьев. Формализуются процессы реализации подзапросов и соединений, соответствующих дереву исходного запроса. А также описывается постановка эксперимента в виртуальном кластере, на примере запроса соединения трех таблиц анализ полученного результата.

В приложении представлен пример текста программы драйвера, реализующего запрос с использованием предложенного метода.

ЗАКЛЮЧЕНИЕ

В магистерской диссертации проведен обзор предметной области и описана структура параллельных систем.

Проведен тщательный анализ существующих подходов к параллельной обработке данных, а также рассматривается технология параллельной распределенной обработки данных MapReduce, ее основные шаги: Map, Shuffle и Reduce. Исследованы преимущества и проблемы реализации этой технологии.

Проведен критический анализ существующих методик реализации соединения таблиц измерений и фактов. Выполнен анализ математического инструментария для моделирования процессов доступа к базам данных, проанализированы недостатки.

Исследованы наиболее часто используемые схемы баз данных. Показано, что наиболее эффективны методы SBJ и SBFCJ доступа к базам данных, реализованные в среде Spark, имеют существенные недостатки.

В качестве метода представления SQL-запроса был предложен вид вложенных деревьев, при котором в каждое дерево входят несколько таблиц измерений и таблица фактов.

Предложена модификация метода с использованием фильтра Блума для выполнения запросов к распределенному хранилищу данных в среде MapReduce/Spark.

Предложен алгоритм автоматической генерации программы для реализации запроса с помощью метода использующего фильтр Блума.

На основании предложенных методов и алгоритмов проведен эксперимент в виртуальном кластере, на примере запроса с соединением 3-х таблиц.

Результаты эксперимента доказали преимущество модифицированного алгоритма с использованием фильтра Блума.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ

1. Гобрик, О. Д. Новые подходы оптимизации запросов в базах данных / Гобрик О. Д., Гуринович А. Б. // Информационные технологии и системы 2019 (ИТС 2019) = Information Technologies and Systems 2019 (ITS 2019) : материалы международной научной конференции, Минск, 30 октября 2019 г. / Белорусский государственный университет информатики и радиоэлектроники; редкол. : Л. Ю. Шилин [и др.]. – Минск, 2019. – С. 300 – 301.

2. Гобрик, О. Д. Сравнительный обзор технологий анализа данных в параллельных системах / Гобрик О. Д. // Информационные технологии и управление: материалы 56-й научной конференции аспирантов, магистрантов и студентов, Минск, 21-24 апреля 2020 года / Белорусский государственный университет информатики и радиоэлектроники; редкол.: Л. Ю. Шилин [и др.]. – Минск, 2020. – С. 73-74.

3. Гобрик, О. Д. Оптимизация обработки данных в параллельных системах / Гобрик О. Д. // Информационные технологии и управление: материалы 57-й научной конференции аспирантов, магистрантов и студентов, Минск, 19-23 апреля 2021 года / Белорусский государственный университет информатики и радиоэлектроники; редкол.: Л. Ю. Шилин [и др.]. – Минск, 2021. – С. 91-92.