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

УДК_____

Геврасёва
Ирина, Петровна

Методы и модели обработки информации в хранилищах данных

АВТОРЕФЕРАТ
на соискание степени магистра технических наук

по специальности 1-40 80 02 Системный анализ, управление и обработка информации

Научный руководитель
Навроцкий Анатолий Александрович
кандидат физико-математических наук, доцент

Минск 2020

ВВЕДЕНИЕ

В настоящее время информация представляет собой один из важнейших ресурсов человеческого общества. Мы живем в мире, который не стоит на месте и постоянно развивается. С появлением интернета объем информации увеличился во много раз и продолжает увеличивается ежедневно, также возрастает и скорость передачи и получения информации. Информационные ресурсы используются практически в любой сфере деятельности человека. Для принятия правильно управленческого решения необходимо иметь релевантную и точную информацию. Ее получение иногда занимает слишком много времени, поэтому возникает необходимость в выборе способа хранения и обработки информации. С целью избежание проблемы анализа накопленных данных предприятия используют системы поддержки принятия решений, основанные на использовании хранилищ данных.

Хранилище данных (Data Warehouse) – это технология, объединяющая структурированные данные из одного или нескольких источников, с целью их дальнейшего анализа для повышения эффективности бизнес-аналитики.

Можно выделить следующие задачи хранилища данных:

– улучшение качества данных;

– подготовка данных для систем поддержки принятия решений;

– интеграция данных из множества источников;

– предоставление доступа к историческим данным;

– минимизация количества несовместимых отчетов.

Качественно спроектированное хранилища данных позволит повысить эффективность использования корпоративных данных для планирования и прогнозирования, анализа и принятия управленческих решений. Тем самым при сокращении времени на данные виды деятельности, у руководителей будет больше времени на другие, что эффективно скажется на их работе, и работе предприятия в целом.

Объектом исследований является хранилище данных. Предметом исследования будут методы и модели обработки информации.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования. Целью данной диссертации является исследование современных методов и моделей обработки информации в хранилищах данных для повышения производительности работы с данными.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Провести анализ и исследование существующих способов хранения данных, а также методов интеллектуального анализа на основе хранилищ данных.

2. Проанализировать модели существующих архитектур хранилищ данных.

3. Провести исследования повышения производительности работы с данными в хранилищах данных.

Новизна полученных результатов.

1. Предложен комбинированный метод построения хранилища данных, совмещающий в себе централизованное хранилище из подхода Билла Инмона и витрины данных из подхода Ральфа Кимболла.

2. Для предложенной системы разработан алгоритм процессов извлечения, трансформации и загрузки данных в промежуточную область, хранилище и витрины данных. Предложенная схема алгоритма обеспечивает высокую эффективность работы хранилища.

Положения, выносимые на защиту.

1. Структурная схема комбинированного хранилища данных, которое совмещает в себе преимущества основных существующих подходов к построению хранилища данных.

2. Схема алгоритма процессов извлечения, трансформации и загрузки данных в промежуточную область, хранилище и витрины данных, позволяющая обеспечить высокую производительность системы.

Апробация результатов диссертации. Основные результаты работы докладывались на 56-й научной конференции аспирантов, магистрантов и студентов учреждения образования «Белорусский государственный университет информатики и радиоэлектроники.

Опубликованность результатов исследования. По теме диссертации опубликовано 4 статьи.

Структура и объем диссертации. Диссертация состоит из содержания, перечня условных обозначений и терминов, введения, четырех глав, заключения, списка использованных источников и приложений. Работа изложена на 58 страницах и содержит: 12 рисунков, список использованных источников из 28 наименований, список 4 публикаций автора и 2 приложения.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Магистерская диссертация представлена в виде пояснительной записки на 58 страницах, состоящей из введения, четырех разделов и заключения.

В первой главе магистерской диссертации произведен анализ предметной области, рассматриваются реляционная, пространственная и многомерная модели данных, их преимущества и недостатки.

Вторая глава посвящена обзору методов интеллектуального анализа данных, применяемых в современных хранилищах данных. Рассмотрены основные классификации стадий, технологических методов и задач ИАД.

В третьей главе рассмотрены основные подходы к построению хранилищ данных, их преимущества и недостатки. А также предложен усовершенствованный комбинированный метод построения хранилища данных, совмещающий в себе преимущества основных подходов.

В четвертой главе рассматриваются алгоритмы повышения производительности работы с данными. Представлен алгоритм процессов извлечения, трансформации и загрузки данных в промежуточную область, хранилище и витрины данных, позволяющая обеспечить высокую производительность системы. На примере конкретного запроса было рассмотрено влияние индексов на производительность выполнения запроса на одинаковых данных.

ЗАКЛЮЧЕНИЕ

В диссертации был проведен анализ существующих способов хранения информации, в рамках которого были рассмотрены существующие модели данных, используемые для организации данных как в базах, так и в хранилищах данных. Каждая из рассмотренных моделей имеет свои особенности, которые определяют области ее применения. Установлено, что реляционную модель данных лучше использовать в базах данных, ориентированных на обработку повседневной транзакционной информации, в то время как, пространственную и многомерную модели – для обработки и анализа историчной информации.

Проведенный обзор показывает, что применение интеллектуального анализа над данными, представленными с помощью систем OLAP в виде информационного гиперкуба, является достаточно эффективным и интегрированным в единую информационно-аналитическую систему. Модели интеллектуального анализа данных могут применяться в конкретных бизнес-сценариях, а именно: прогнозирование, риск и вероятность, рекомендации, поиск последовательностей и группировка.

Проанализировав особенности основных подходов к организации хранилищ данных, позволяющих производить анализ информации для принятия дальнейших решений, была предложена архитектура хранилища, совмещающая в себе централизованное хранилище данных из подхода Билла Инмона, спроектированное с использованием пространственного метода представления данных, и витрины данных из подхода Ральфа Кимболла.

Было проанализировано влияние наличия или отсутствия индексов на выполнение запросов в хранилищах данных. Установлено, что запрос, выполненный над таблицами, имеющими первичные и вторичные ключи, занимает примерно на 25% меньше времени, чем над такими же таблицами, но без индексов, что позволяет снизить временные потери при генерации отчетов за счет увеличения производительности выполнения запросов.

Проведенное исследование показало, что на эффективность работы хранилища данных влияют ETL-процессы, которые являются не только инструментом переноса данных из источников в хранилища, но и подготовки данных к анализу. Был создан обобщенный алгоритм извлечения, преобразования и загрузки данных в промежуточную область, централизованное хранилище и витрины данных. Предложенных алгоритм обеспечивает высокую эффективность работы хранилища.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

[1] Геврасёва, И. П. Подходы к построению хранилищ данных / И. П. Геврасёва // 56-я научная конференция аспирантов, магистрантов и студентов учреждения образования «Белорусский государственный университет информатики и радиоэлектроники»: материалы конференции по направлению 2: Информационные технологии и управление, Минск, 21–24 апреля 2020 г. / редкол.: Л. Ю. Шилин [и др.]. – Минск: БГУИР, 2020. – С. 67.

[2] Геврасёва, И. П. Модель многомерного представления данных в хранилищах данных / Геврасёва И. П. // 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. 302 – 303.

[3] Геврасёва, И. П. Обработка информации в хранилищах данных / И. П. Геврасёва // 55-я юбилейная научная конференция аспирантов, магистрантов и студентов учреждения образования «Белорусский государственный университет информатики и радиоэлектроники»: материалы конференции по направлению 2: Информационные технологии и управление, Минск, 22–26 апреля 2019 г. / редкол.: Л. Ю. Шилин [и др.]. – Минск: БГУИР, 2019. – С. 73.

[4] Heurasiova, I. P. Data warehouse modeling / I. P. Heurasiova // Проблемы экономики и информационных технологий: сборник тезисов докладов 55-й юбилейной научной конференции аспирантов, магистрантов и студентов, Минск, 22 – 26 апреля 2019 г. / Белорусский государственный университет информатики и радиоэлектроники. – Минск, 2019. – С. 185 – 189.