УНИВЕРСИТЕТ ИТМО

Технологии хранения больших данных

Лекция №5. Поисковые индексы и эффективное хранение и применение индексов на диске – хэш-таблицы, деревья поиска, пространственные индексы, полнотекстовый поиск

Ирина Алексеевна Радченко
iradche@gmail.com

Введение

Обзор темы

- Значение поисковых индексов
- Виды индексов: хэш-таблицы, деревья поиска, пространственные индексы, полнотекстовый поиск

Цели лекции

- Понимание различных типов индексов
- Изучение методов хранения и применения индексов на диске

Поисковые индексы

Определение и значение

- Ускорение поиска данных
- Снижение времени доступа к данным

Основные виды индексов

- Первичные и вторичные индексы
- Уникальные и неуникальные индексы

Хэш-таблицы

Принципы работы

- Ассоциативное хранение данных
- Использование хэш-функций для вычисления индексов

Преимущества и недостатки

- Высокая скорость доступа
- Ограничения по размеру и коллизии

Пример хэш-таблицы

Структура хэш-таблицы

- Ключи и значения

Пример хэш-функции

```
def hash_function(key):
    return sum(ord(char) for char in key) %
size
```

Деревья поиска (B-деревья)

Принципы работы

- Дерево с балансировкой для эффективного поиска
- Узлы с множеством ключей и дочерних узлов

Преимущества и недостатки

- Эффективность поиска, вставки и удаления
- Сложность реализации

Пример B-дерева

Структура B-дерева

- Узлы, ключи и дочерние узлы

Пример поиска

```
SELECT * FROM BTree WHERE key = 'value';
```

Пространственные индексы (R-деревья)

Принципы работы

- Иерархическая структура для пространственных данных
- Узлы представляют собой многоугольники

Преимущества и недостатки

- Эффективность работы с геоданными
- Сложность обновления

Пример R-дерева

Структура R-дерева

- Узлы и их геометрические области

Пример запроса

```
•SELECT * FROM SpatialIndex WHERE region = 'specified_region';
```

Полнотекстовый поиск

Принципы работы

- Индексирование текста для быстрого поиска
- Использование токенизации и лемматизации

Преимущества и недостатки

- Быстрый поиск по тексту
- Требования к пространству хранения

Пример полнотекстового поиска

Структура полнотекстового индекса

- Индексные файлы и токены

Пример запроса

```
SELECT * FROM Documents WHERE MATCH(content) AGAINST ('search_term');
```

Сравнение индексов

Хэш-таблицы vs. Деревья поиска

- Скорость доступа, эффективность обновления

Пространственные индексы vs. Полнотекстовый поиск

- Специализированные задачи, сложность реализации

Хранение индексов на диске

Методы хранения

- Последовательное хранение, файловые структуры

Оптимизация доступа

- Использование буферов и кэширования

Пример хранения на диске

Структура хранения B-дерева

- Узлы на диске, использование блоков

Пример организации хранения

```
# Команды для хранения индексов
CREATE INDEX btree_index ON table (column);
```

Эффективное использование индексов

Оптимизация запросов

- Использование индексов для ускорения поиска

Поддержка индексов

- Обновление и удаление индексов

Практические примеры и кейсы

Реальные примеры использования различных типов индексов

- Транзакционные системы, системы поиска, геоинформационные системы

Заключение

Подведение итогов

- Важность выбора подходящего типа индекса для конкретной задачи
- Влияние индексов на производительность систем

Дополнительные материалы и литература

УНИВЕРСИТЕТ ИТМО

Вопросы и обсуждение

Вопросы от студентов

Обсуждение практических примеров

Дальнейшие шаги и изучение

ITMO more than a UNIVERSITY

Благодарность за внимание

Контактная информация

- Email: iradche@gmail.com
- Телеграм: @dadaistka

Следующая лекция

- Тема следующей лекции:

  «Форматы хранения данных – таблицы, protobuf, avro»

Спасибо за внимание!

www.ifmo.ru

ITMO more than a UNIVERSITY