ПОЛИТЕХ
Санкт-Петербургский
Политехнический Университет
Петра Великого

Санкт-Петербургский политехнический университет Петра Великого
Институт компьютерных наук и технологий
Высшая школа программной инженерии

Системы анализа больших данных (САБД)

3 семестр

Направление: *09.04.04 – «Программная инженерия»*

Преподаватели:
Ковалев Артем Дмитриевич
Никифоров Игорь Валерьевич, к.т.н.

Задание на семестр

- Написание курсового проекта по теме:
  - Исследование подходов к созданию высокопроизводительного, масштабируемого сервиса для дедупликации данных в хранилище
- Консультации
- В конце семестра — защита курсового проекта

Цель курсового проекта

- Цель работы состоит в реализации системы оптимального хранения данных за счет использования подхода дедупликации данных и проведении тестирования для измерения производительности созданного прототипа.

Задачи выполняемой работы

- изучить подходы к оптимизации хранения данных в традиционных базах данных;
- выбрать стек технологий, необходимый для создания прототипа системы дедупликации;
- разработать систему оптимального хранения данных на диске;
- провести нагрузочное тестирование и установить зависимость скорости чтения/записи данных в локальное хранилище в зависимости от размера сегмента;
- установить оптимальный размер сегмента блока данных, при котором наблюдается:
  - максимальная скорость записи
  - максимальная скорость чтения
  - процент ошибки восстановления данных в зависимости от использованного алгоритма hash

Краткое описание курсового проекта

- входной поток данных разделяется на блоки (сегменты) заданного объема (4 байта)
- для каждого блока вычисляется hash
- значение hash проверяется с таблицей уже существующих hash для данных находящихся на носителе
- если значение hash для данных уже есть, то для входного потока сохраняется лишь ссылка на уже существующие данные
- если данных нет, то новый блок сохраняется на диске и hash этого блока записывается в таблицу с соответствующим расположением

Рекомендации к выполнению

- Источник данных
  - любой
- Хранилище данных
  - локальный диск
- Хранилище таблиц hash-значений и ссылок
  - MongoDB/PostgreSQL/Redis/HBase/etc
- Алгоритм вычисления hash
  - MD5, SHA128, SHA256, SHA512, etc

Отчет

- демонстрация программной реализации системы хранения дедуплицированных данных
- отчет с полным описанием разработанной системы

Спасибо за внимание!

Вопросы?