Обзор алгоритмов content-defined chunking(CDC)

Системы хранения данных бывают разных типов, в частности, это могут быть хранилища резервных копий. Подобные хранилища отличаются высокой плотностью хранения данных за счет сжатия и дедупликации. Основная идея дедупликации заключается в том, что входной набор данных разделяется системой на блоки и для каждого блока вычисляется подпись. Блок с уникальной подписью сохраняется в базу данных. Если в системе уже имеется некоторая подпись, то новый блок с такой подписью не сохраняется. Одним из способов деления на блоки является метод content-defined chunking, заключающийся в делении данных на блоки по срабатыванию специального хеша, который с высокой вероятностью даст одинаковые границы на одинаковых данных.

Основными целями работы являются:

- выполнение обзора основных алгоритмов content-defined chunking;
- проведение сравнения рассмотренных алгоритмов по объему используемой памяти, скорости работы на одном ядре и коэффициенту дедупликации;
- исследование возможности оптимизации алгоритмов за счет применения векторизации и/или распараллеливания;
- разработка оптимизированного алгоритма и анализ его производительности.

Планируемые результаты

- Теоретический обзора метода CDC и основных алгоритмов;
- Реализация основного пула алгоритмов CDC;
- постановка и проведение экспериментального сравнения алгоритмов CDC по производительности, использованию памяти, коэффициенту дедупликации на датасетах enron2, vmvare image, ubuntu iso, linux kernel source code;
- анализ экспериментальных результатов;
- выявление векторизуемых или параллелизируемых алгоритмов;
- разработка, реализация и анализ эффективности оптимизированного с помощью распараллеливания и/или векторизации алгоритма CDC.

Этапы выполнения проекта

1) Знакомство с предметной областью: дедупликация, CDC. Выполнение обзора основных алгоритмов CDC
2) Реализация алгоритмов, подготовка экспериментального стенда
3) Проведение экспериментального сравнения алгоритмов
4) Анализ полученных результатов
5) Исследование возможности применения распараллеливания и/или векторизации для рассмотренных алгоритмов. Разработка оптимизированных алгоритмов, их реализация, и сравнение с базовыми версиями.

Технологии

Предпочтительный язык выполнения проекта — Rust (возможно использование Rust в комбинации с C для использования intrinsic).

Некоторые ссылки

- Общие концепции и идеи CDC
- Пример алгоритма для рассмотрения
- Об асинхронном программировании в Rust

Пожелания к участникам

- Базовое владение Rust или заинтересованность в таковом
- Заинтересованность в теме