Frequently Based Chunking
обзор алгоритмов

Цели и задачи

Цель:

1) Анализ алгоритмов FBC

Задачи:

1) Сравнение реализаций алгоритмов FBC
2) Анализ возможностей оптимизации
3) Разработка алгоритмов и анализ их произовдительности

Особенности подхода

1) Выделение новых чанков из уже существующих
2) Зависимость от алгоритма 1 стадии (CDC)
3) Анализ частоты

Выделение новых чанков

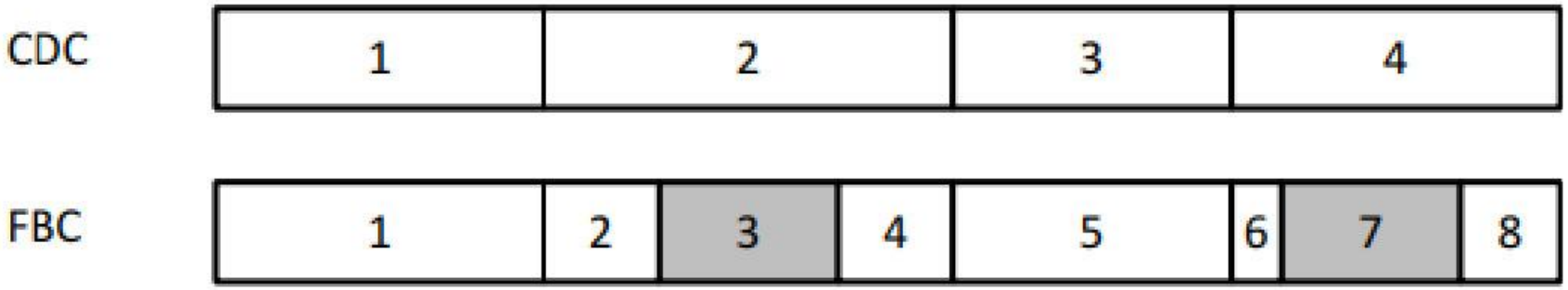

Диаграмма 1.1 - Выделение новых чанков

источник [1]

2 стадия (post-CDC)

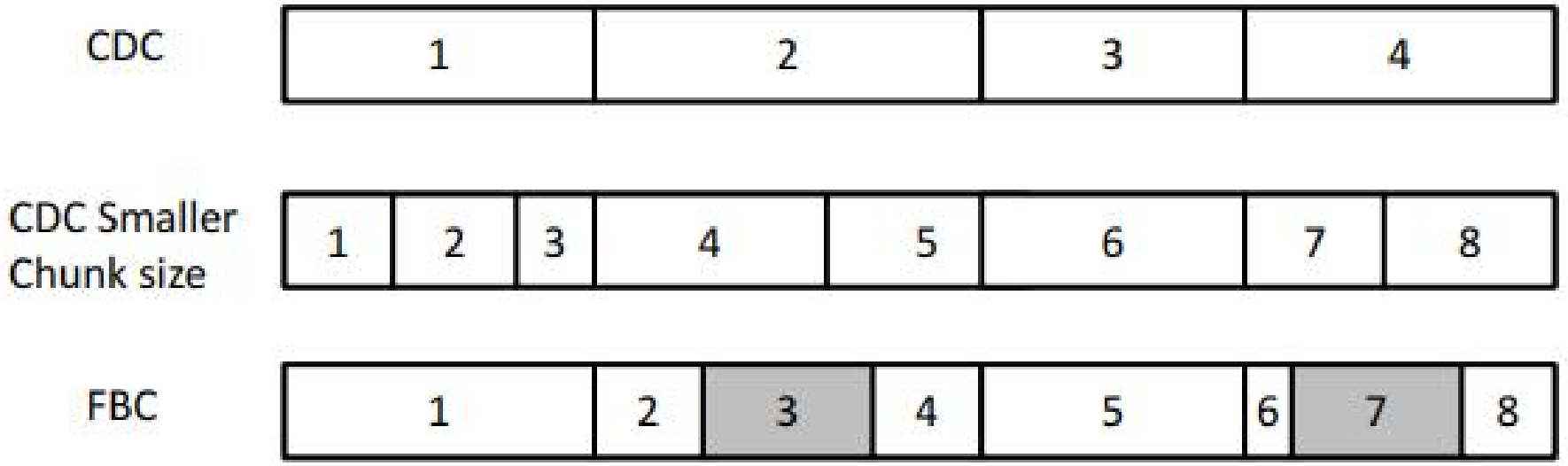

Диаграмма 1.2 - Зависимость от выбора алгоритма CDC

источник [1]

Анализ частоты

1) На основе конкретного датасета [2]
2) Предварительная фильтрация пересечений
3) Частотный анализ

Реализация

Используемые языки:

Rust 1.77.0

Возможно ипользование при прототипировании:

Python, C (различных версий)

План работы

1) Анализ алгоритмов (практически завершен)
2) Прототипирование (начато)
3) Разработка
4) Тестирование
5) Интеграция
6) Сравнение

Источники / примечания

[1] - Lu, G.; Jin, Y.; Du, D.H. Frequency based chunking for data de-duplication. In Proceedings of the 2010
IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication
Systems, Miami Beach, FL, USA, 17–19 August 2010; IEEE: New York, NY, USA, 2010.
[2] Saeed, A.S.M.; George, L.E. Data deduplication system based on content-defined chunking using bytes pair frequency occurrence.Symmetry 2020, 12, 1841