Обзор литературы (набросок)

Building a High-performance Deduplication System

(GuoEfstathopoulos.pdf)

Вводятся три механизма, призванных улучшить эффективность дедупликации:

1. Прогрессивное выборочное индексирование (*Progressive sampled indexing*). Индекс — сопоставление идентификатора фрагмента (хеша) и расположения его полного содержимого. Вместо единого индекса, содержащего информацию о всех блоках, предлагается хранить информацию, актуальную для конкретного файла, вместе с остальными метаданными файла (в списке его блоков). В таком случае вместо полного глобального индекса можно использовать выборочный, содержащий информацию только о некоторых “горячих” фрагментах. Предлагаются некоторые методы спекулятивного принятия решений о взятии фрагментов в индекс. Прогрессивность относится к динамическому выбору размера “горячего” индекса (*sampling rate*) в обратной пропорции к объёму хранимых данных.
2. Сгруппированная пометка-и-уборка (*Grouped mark-and-sweep*). Вариация классического алгоритма сборки мусора (ставших ненужными блоков данных), работающая не над всеми файлами во всех бекапах, а по возможности ограничивающаяся только затронутыми группами. Шаг пометки проводится только для изменённых групп, шаг уборки — только для используемых в них контейнеров.
3. Многопоточная модель взаимодействия клиент-сервер на основе событий (*Event-driven multi-threaded client-server interaction model*). Предложен асинхронных протокол на основе событий. Вычисления по возможности распределяются, тогда как запросы группируются. Адаптация к меняющимся нагрузкам может осуществляться с помощью вариации размеров различных очередей.

Благодаря предложенным механизмам разработанный прототип продемонстрировал отличную масштабируемость, высокую пропускную способность и низкую деградацию эффективности дедупликации.

Demystifying Data Deduplication

(mandagere2008.pdf)

Обозреваются таксономия понятий в дедупликации и численные параметры, сопутствующие различным вариантам резервного копирования, в реальных условиях.

Выделяются три свойства решений в дедупликации:

1. Размещение функционала дедупликации
    1. Клиент-сервер резервного копирования.
    2. Дедуплицирующая аппаратура.
    3. Массив хранения данных.
2. Время проведения дедупликации
    1. Синхронная дедупликация (*In-Band*).
    2. Асинхронная дедупликация (*Out-of-band*).
3. Алгоритм поиска повторов
    1. Дельта-кодирование.
    2. Хеширование целых файлов.
    3. Хеширование блоков фиксированного размера.

4. Хеширование блоков переменного размера.

A Study of Practical Deduplication

(meyer2012.pdf)

Для персональных компьютеров была проведена оценка сравнительной эффективности различных методов дедупликации. Было получено, что дедупликация на уровне целых файлов достигает 75% экономии места, обеспечиваемой наиболее агрессивной дедупликацией на уровне отдельных блоков, при использовании в активных файловых системах, и 87% при использовании в резервных копиях. Также было получено, что распределение размеров файлов смещается в направлении крупных неструктурированных файлов, а фрагментация данных на диске на практике несущественна.

Data Deduplication techniques

(qinluhe2010.pdf)

Работа, при более детальном изучении, не заслуживает рассмотрения в НИР.

A Comprehensive Study of thePast, Present, and Future of Data Deduplication

(xia2016.pdf)

Рассмотрена таксономия понятий в дедупликации и достижения, описанные в других работах, а также перспективные открытые проблемы.

Приведена статистика степени сжатия данных за счёт дедупликации в различных практических контекстах — на персональных компьютерах и на серверах, при блочной или полнофайловой дедупликации, и других параметрах — полученная на основе 5 других работ. Рассмотрены механизмы разбиения потока данных на блоки, основанные на алгоритме Рабина, с различными доработками, предложенными в 13 других работах. Также рассмотрены способы вычислительной оптимизации задач дедупликации, предложенные в 7 других работах. Кроме того, рассмотрены способы индексирования известных блоков и поиска сходств / дельта-кодирования. Проведён краткий обзор подходов ко внедрению криптографии в системы с дедупликацией. Рассмотрены сценарии, где дедупликация может приносить практическую пользу.

Статья заслуживает большого внимания в НИР.

A Study on Deduplication Techniques over Encrypted Data

(akhila2016.pdf)

Проведён обзор методов использования шифрования совместно с дедупликацией, предложенных в других работах. Можно выделить принципиальные методы:

- Шифрование, завязанное на сообщение (*message-locked encryption*) / конвергентное шифрование (*convergent encryption*). Ключ шифрования зависит только от содержимого шифруемого сообщения. Позволяет сравнивать зашифрованные сообщения точно так же, как и исходные.
- Доказательство владения (*proof of ownership*). Интерактивный алгоритм, подтверждающий владение полными данными, соответствующими некоторому хешу. Необходимо для противодействия атакам внедрения фиктивных данных, якобы дублирующих некоторый секретный блок.

- Генерация ключей с участием сервера. Многие методы полагаются на доверенную третью сторону, участвующую в генерации ключей. Это позволяет реализовать конвергентное шифрование, не допуская при этом атак по раскрытию открытого текста из узкого спектра.
- Хранение “популярных” данных без шифрования. Предполагается, что данные с высокой степенью дупликации не носят секретный характер и могут храниться без шифрования.

DupLESS: Server-Aided Encryption for Deduplicated Storage

(sec13-paper_bellare.pdf)

Предлагается способ усиления шифрования, завязанного на сообщения, призванный устранить возможность атак, раскрывающих открытый текст из узкого спектра. При этом используется сервер ключей. Предоставлено описание криптографического протокола и его реализация, работающая с существующими крупными облачными хранилищами без какой-либо особой поддержки с их стороны. Разработанный протокол не приводит к значительным потерям эффективности дедупликации. Безопасность протокола достигается за счёт использования дополнительного секретного параметра при генерации ключей, а также использования протокола забывчивой передачи, чтобы избежать разглашения этого параметра.

Encrypted Data Management with Deduplication in Cloud Computing

(yan2016-2.pdf)

Предлагается способ использования шифрования с дедупликацией за счёт шифрования на основе аттрибутов (*attribute-based encryption*).

[TODO]

Deduplication on Encrypted Big Data in Cloud

(yan2016.pdf)

Предлагается способ использования шифрования с дедупликацией за счёт проверки владения и опосредованного пере-шифрования (*ownership challenge and proxy re-encryption*).

[TODO]