



Google File System / HDFS

Ильнур Шугаев

Google File System (GFS)¹

Цели

- ✓ Производительность
- ✓ Масштабируемость
- ✓ Надежность
- ✓ Доступность

¹Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. "The Google file system". In: (2003).

Table of Contents

1. Introduction

2. Дизайн

Наблюдения и Предположения
Архитектура

3. Master

4. Надежность и средства диагностики

5. HDFS

Наблюдения

1

Отказ компонент
— норма, а не
исключение

2

Поддержка большого
количества маленьких
файлов — сложно

3

Мутации —
дописывание в
конец

Предположения

- Система состоит из большого числа компонент, которые могут часто отказывать
- Система хранит преимущественно большие файлы (> 100 MB)
- Основные операции: потоковое чтение, запись в конец
- Много клиентов могут одновременно делать запись в конец файла

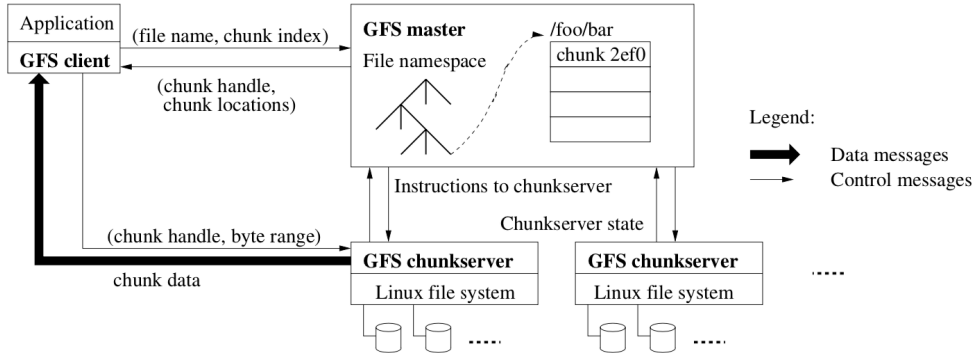


Figure: Архитектура системы

Хранение файлов

- 1 Файл разбивается на куски (chunks) фиксированного размера (64 MB)
- 2 Кусок идентифицируется уникальным `chunk_handle`, который выдается мастером
- 3 Куски файла раскидываются и реплицируются по разным chunkserver'ам, которые выбирает мастер

Master

- Хранит все метаданные
- Создает/реплицирует чанки
- Сборка мусора
- Общается с chunkserver'ами с помощью HeartBeat сообщений
- Обрабатывает запросы связанные с метаданными
- Ведет лог всех операций



Chunkserver

- Хранит данные на локальном диске
- Обменивается данными с клиентом напрямую
- Обменивается данными с другими chunkserver'ами

Чтение файла

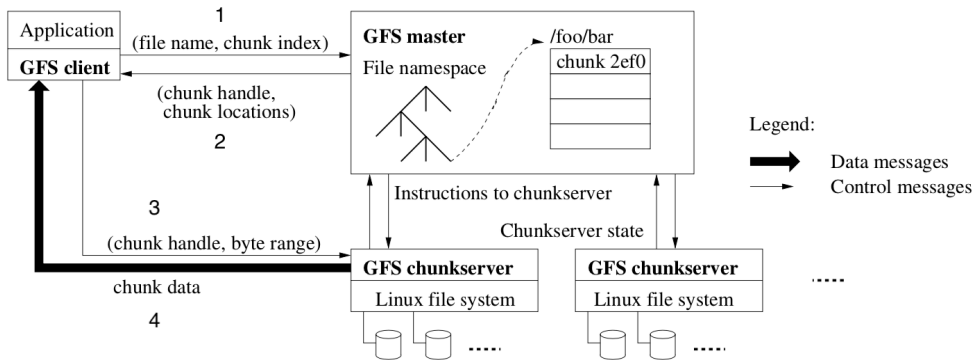


Figure: Этапы чтения файла

Table of Contents

1. Introduction

2. Дизайн

3. Master

Создание, ре-репликация, балансировка

Сборка мусора

4. Надежность и средства диагностики

5. HDFS

Создание куска

Факторы влияющие на то, на каком chunkserver'e будет создан кусок

- ✓ Утилизация диска
- ✓ Утилизация сети
- ✓ Как давно был создан последний кусок
- ✓ Где находится сервер

Ре-репликация

1. Кусок ре-реплицируется, как только число реплик становится ниже заданного уровня
2. Ре-репликация происходит по приоритету

Балансировка

Мастер отдает команды на перераспределение кусков для лучшей утилизации дисков и сети

Механизм удаления файлов

1. Файл помечается удаленным (память не освобождается)
2. Освобождение ресурсов происходит во время очередного цикла сборки мусора (файл должен числиться удаленным больше определенного периода)

Table of Contents

1. Introduction

2. Дизайн

3. Master

4. Надежность и средства диагностики

Надежность

Средства диагностики

5. HDFS

Стратегии

- Быстрое восстановления мастера и chunkserver'ов
- Репликация данных
- Репликация состояния мастера

Логгирование

Изменения становятся доступны клиентам только после того как все действия будут внесены в лог

Table of Contents

1. Introduction

2. Дизайн

3. Master

4. Надежность и средства диагностики

5. HDFS




Hadoop Distributed File System (HDFS)²³

Отличия от GFS

- Open-source реализация GFS
- Master — NameNode
- Chunkserver — DataNode

²Konstantin Shvachko et al. "The hadoop distributed file system.". In: *MSS7*. vol. 10. 2010, pp. 1–10.

³Tom White. *Hadoop: The definitive guide*. " O'Reilly Media, Inc.", 2012.

-  Ghemawat, Sanjay, Howard Gobioff, and Shun-Tak Leung. "The Google file system". In: (2003).
-  Shvachko, Konstantin et al. "The hadoop distributed file system.". In: *MSST*. Vol. 10. 2010, pp. 1–10.
-  White, Tom. *Hadoop: The definitive guide*. " O'Reilly Media, Inc.", 2012.



Вопросы?

Ильнур Шугаев