

FS Dedupe

Владислав Белогрудов, EMC
vlad.belogrudov@gmail.com

Проблема

Одинаковые или похожие файлы -
версии документов, фотографии,
архивы

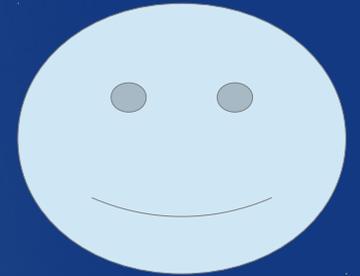


Сложно отыскать
нужный файл

Неэффективный
расход носителей



Есть

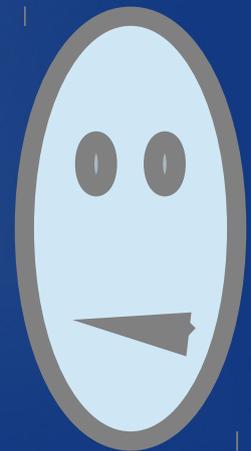


Прототип системы

- C++ backend, Java frontend
- Кроссплатформенность
- фильтрация по размеру и типу файлов
- LCS для определения похожести
- MD4/5 для нахождения одинаковых файлов

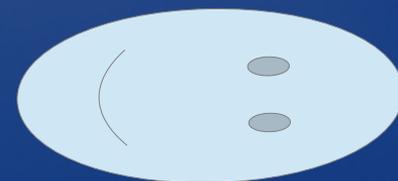
Задачи

- Добавление алгоритмов и стратегий
 - Levenstein
 - архивы
 - графические форматы
 - выбор наилучшего алгоритма “на лету”
- Оптимизация
- Тестирование
- Управление дубликатами



Требования к команде

- C/C++
- Java
- Linux
- Make, GDB, MS VS
- Желание изучить и внедрить подходящие алгоритмы



Спасибо!