百万件くらいのデータの扱い方

Unix大量ゴミファイル事件簿

Masafumi Yokoyama

Rabbitについて

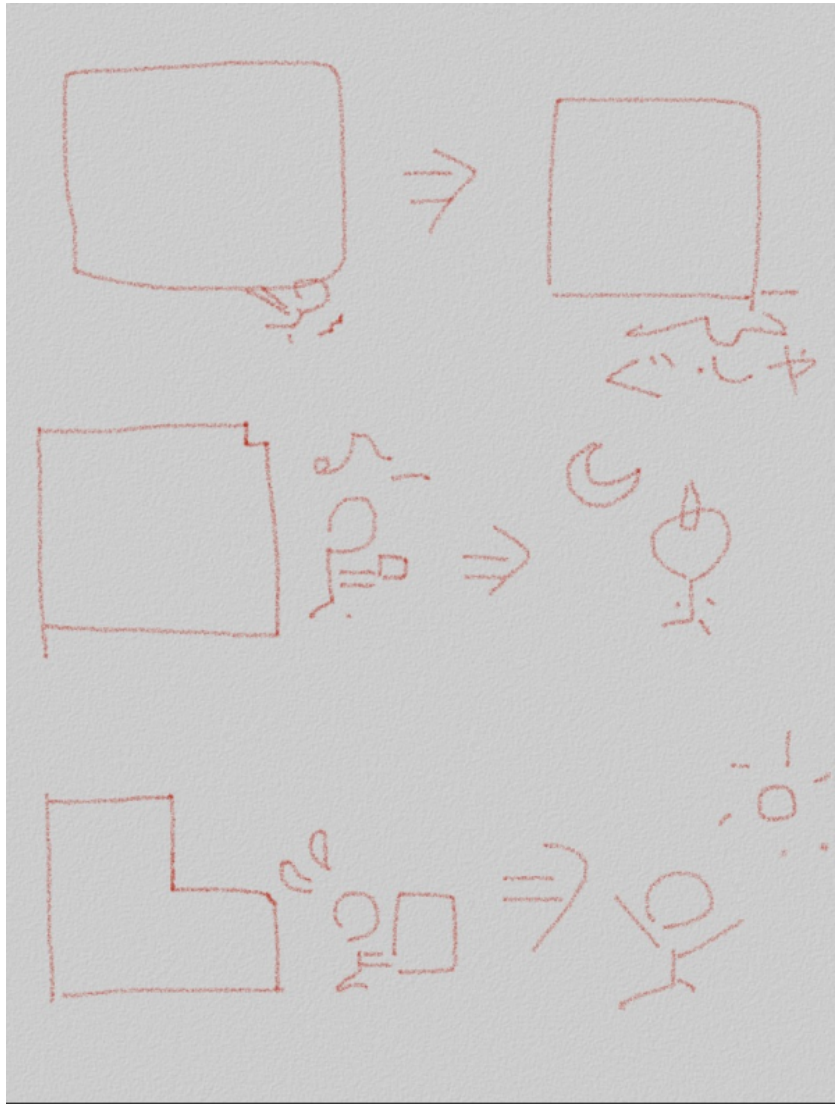
プレゼンテーションツール

- 実装: Ruby/GTK+
- 動作: PC-UNIX/Win/Mac
- 文章とデザインの分離
  - バージョン管理しやすい

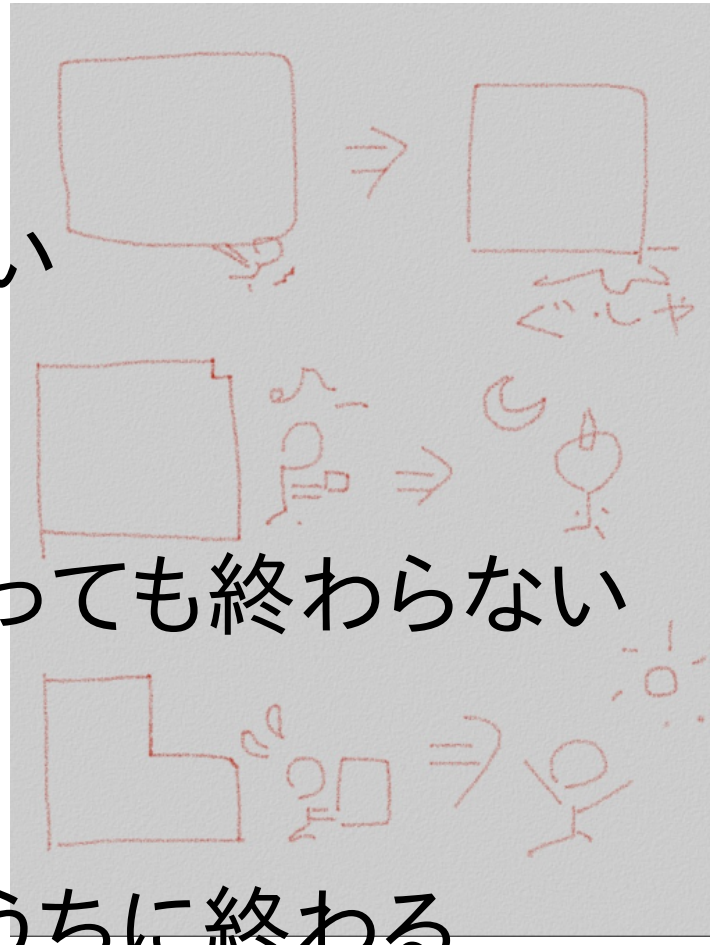
テーマ

- 件数が多いデータの扱い方

キーワードは『荷物運び』



一度に運ぶ量



- 大量
  - 持てない
- 少量
  - 夜になっても終わらない
- 適量
  - 明るいうちに終わる

考え方

- 件数が多いデータは、適量に分けて処理する
  - まずは適量を見極める

具体例

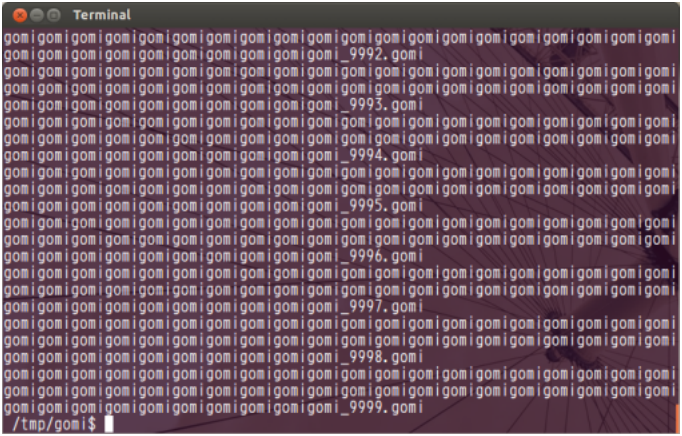
- 大量ファイルの削除
  - Unixサーバー
  - GUIが使えない

問題

- ディレクトリ内の.gomiファイルを全て削除したい

状況分析

ls(dir)コマンドの末端



分析結果

- 名前が長いファイルが大量にある
  - ファイル名: 210バイト
  - ファイル数: 10,000

```
# テストデータの作り方
$ cd /tmp ; mkdir gomi ; cd gomi
$ ruby -e '0.upto((10 ** 4) - 1) {|i| `touch #{"gomi" * 50}_#{"%04d" % i}.gomi` }'
$ (cd .. ; tar czf gomi.tar.gz gomi) #バックアップ
$ (cd .. ; tar xzf gomi.tar.gz)      #復元
```

(1) 普通に削除

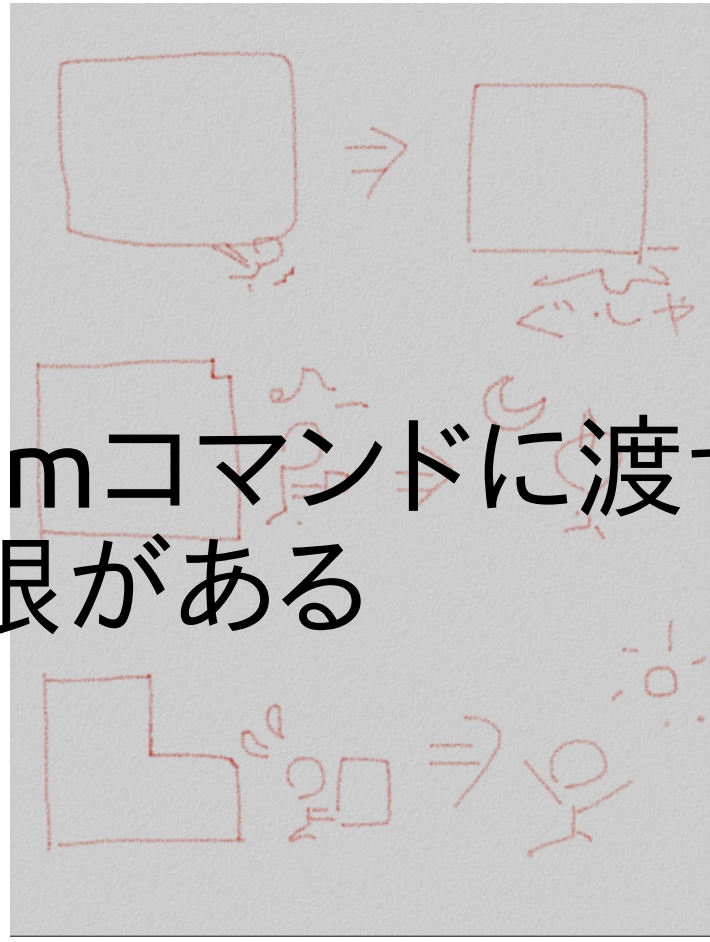
```
$ ls | wc -l  #ファイル数を数える
10000

$ rm *.gomi
/bin/rm: cannot execute [引数リストが長すぎます]

$ ls | wc -l
10000
```

消せない!

原因



- 一度にrmコマンドに渡せる引数には上限がある

NG

(2) 1ファイルずつ削除

```
$ rm gomi...gomi_0000.gomi
$ rm gomi...gomi_0001.gomi
$ rm gomi...gomi_0002.gomi
$ rm gomi...gomi_0003.gomi
$ rm gomi...gomi_0004.gomi
$ rm gomi...gomi_0005.gomi
$ rm gomi...gomi_0006.gomi
$ rm gomi...gomi_0007.gomi
$ rm gomi...gomi_0008.gomi
$ rm gomi...gomi_0009.gomi
$ rm gomi...gomi_0010.gomi
$ rm gomi...gomi_0011.gomi
$ rm gomi...gomi_0012.gomi
$ rm gomi...gomi_0013.gomi
$ rm gomi...gomi_0014.gomi
$ rm gomi...gomi_0015.gomi
$ rm gomi...gomi_0016.gomi
$ rm gomi...gomi_0017.gomi
$ rm gomi...gomi_0018.gomi
$ rm gomi...gomi_0019.gomi
$ rm gomi...gomi_0020
```

徹夜

NG

(3) 適量ずつ削除

```
$ ls | wc -l
10000
$ rm *0.gomi
$ rm *1.gomi
$ rm *2.gomi
$ rm *3.gomi
$ rm *4.gomi
$ rm *5.gomi
$ rm *6.gomi
$ rm *7.gomi
$ rm *8.gomi
$ rm *9.gomi
$ ls | wc -l
0
```

(3) 適量ずつ削除

```
$ find . -name "*.gomi" | wc -l
10000

$ find . -name "*.gomi" | xargs rm

$ find . -name "*.gomi" | wc -l
0
```

OK!

落とし穴

- 以下のやり方だと、rmコマンドが一万回呼ばれる

```
$ find . -name ”*.gomi” | xargs -i rm ’{}’

$ find . -name ”*.gomi” -exec rm ’{}’ ¥;
```

遅い

- コマンドを呼ぶ回数が多いと遅くなる

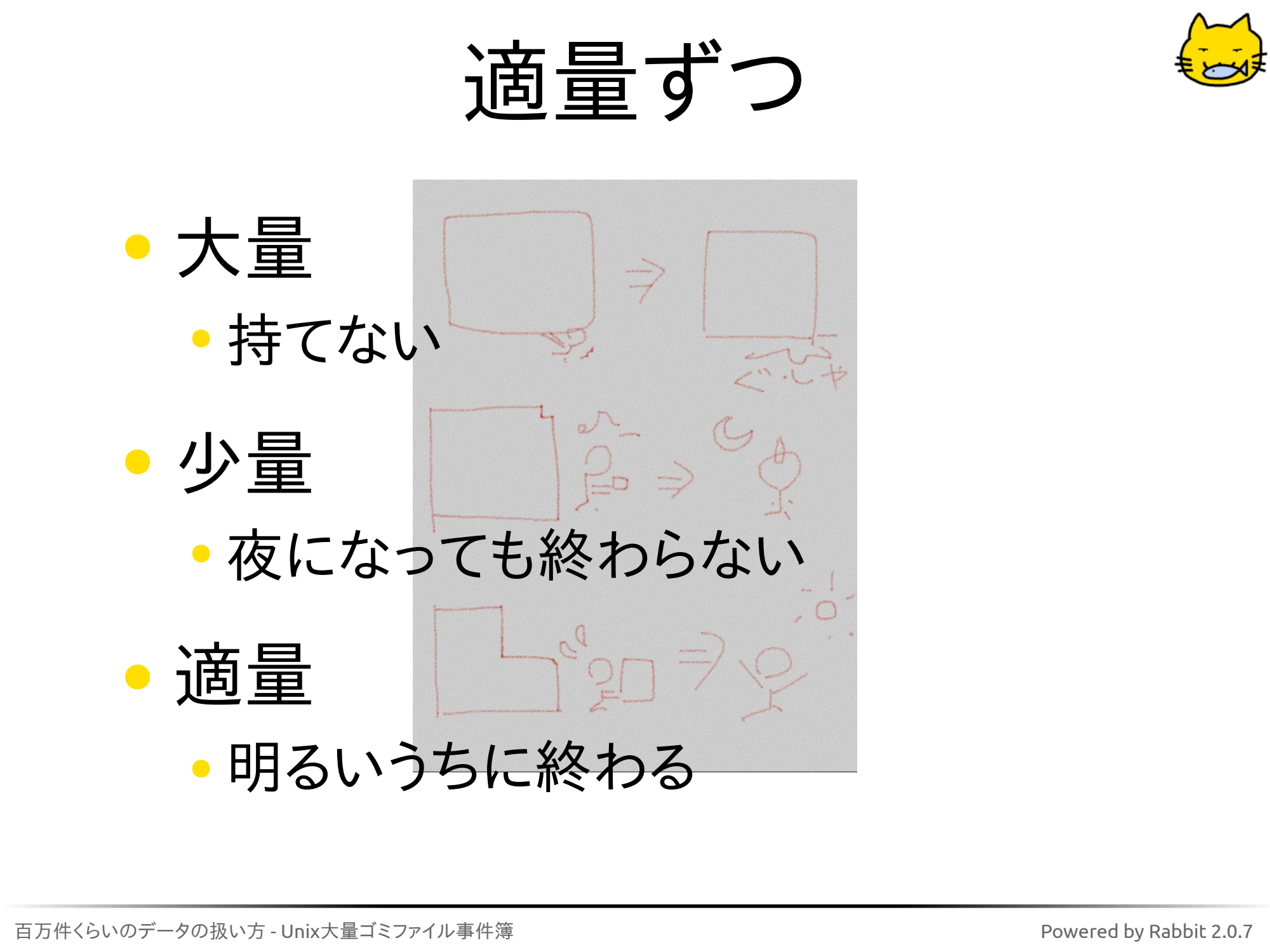
```
$ find . -name "*.gomi" | xargs rm
#=> 0.147秒

$ find . -name "*.gomi" | xargs -i rm '{}'
#=> 14.120秒

$ find . -name "*.gomi" -exec rm '{}' ¥;
#=> 18.512秒
```

徹夜の恐れ

適量ずつ



- 大量
  - 持てない
- 少量
  - 夜になっても終わらない
- 適量
  - 明るいうちに終わる

その他の例

- DBMSのトランザクション処理
  - DB2では100件くらいずつコミットする
  - IMPORTのcommitcountオプション
- プロセス数やスレッド数
  - メモリやCPUが100%で固定されない程度に増やすと速い

おわり