Web サーバのログ解析プログラム P3-2(Perl)

UNIX 的ログ解析 *(*言いすぎ*)*

小山浩之

oyama@cpan.org

Tokyo/Shibuya Perl Mongers

UNIX的ってなによ？

- スモールイズビューティホー！
- 一つのプログラムは一つのことをうまくやらせる
- ソフトウェアのテコを有効に活用する
- すべてのプログラムをフィルタにする
- 部分の総和は全体よりも大きい

UNIX という考え方 (ISBN4-274-06406-9) より。

具体的にどう実装する？

ログ解析の要件を満たすためには？

- ログから"時刻"と"アクセス元"を抽出する
- "時刻"と"アクセス元"のリストから、"セッション"を抽出
- "セッション"の情報をもとに、レポートを出力

具体的にどう実装する？

ログ解析の要件を満たすためには？

- ログから"時刻"と"アクセス元"を抽出する
- "時刻"と"アクセス元"のリストから、"セッション"を抽出
- "セッション"の情報をもとに、レポートを出力

小さく単機能な3つのフィルタプログラムを組み合わせて問題を解決する！→通信にはPIPEをつかう。

3つのフィルタ

- parser
- sorter
- reporter

3つのフィルタ

- parser
  stdin から行単位で読み込み、正規表現で"時刻"と"アクセス元"を抽出。"時刻"は Coordinated Universal Time に変換し、stdout へ。
- sorter
- reporter

3つのフィルタ

- parser
  stdin から行単位で読み込み、正規表現で"時刻"と"アクセス元"を抽出。"時刻"は Coordinated Universal Time に変換し、stdout へ。
- sorter
  stdin から行単位で読み込み、"アクセス元"ごとの"セッション"を抽出。コンテナには Hash を使用する。"アクセス元", "訪問時刻", "退出時刻"を stdout へ。
- reporter

3つのフィルタ

- parser
  stdin から行単位で読み込み、正規表現で"時刻"と"アクセス元"を抽出。"時刻"は Coordinated Universal Time に変換し、stdout へ。
- sorter
  stdin から行単位で読み込み、"アクセス元"ごとの"セッション"を抽出。コンテナには Hash を使用する。"アクセス元", "訪問時刻", "退出時刻"を stdout へ。
- reporter
  stdin から行単位で読み込み、"アクセス元"の"総滞在時間", "訪問回数", "平均滞在時間"を stdout に出力する。

使い方

```
% cat access.log | parser | sorter 300 | \
grep '^xxx.xxx.xxx ' | reporter
Total:     15[sec]
Access: 1
Mean:    15.00[sec]
```

この実装のメリット

- それぞれ、やるべき事が明確なので実装が楽でシンプル。
- 機能追加がしやすい。
- プログラム単位で再利用が可能。

この実装のメリット

- それぞれ、やるべき事が明確なので実装が楽でシンプル。→保守しやすい
- 機能追加がしやすい。
- プログラム単位で再利用が可能。

この実装のメリット

- それぞれ、やるべき事が明確なので実装が楽でシンプル。→保守しやすい
- 機能追加がしやすい。 → ログが時系列じゃ無い？なら sort コマンドを噛ます。
- プログラム単位で再利用が可能。

この実装のメリット

- それぞれ、やるべき事が明確なので実装が楽でシンプル。→保守しやすい
- 機能追加がしやすい。 → ログが時系列じゃ無い？なら sort コマンドを噛ます。
- プログラム単位で再利用が可能。 →プログラミング不要な再利用

この実装のメリット

- それぞれ、やるべき事が明確なので実装が楽でシンプル。→保守しやすい
- 機能追加がしやすい。 → ログが時系列じゃ無い？なら sort コマンドを噛ます。
- プログラム単位で再利用が可能。 →プログラミング不要な再利用