

Groonga 導入事例

佐藤 博之

Groonga Meetup 2015
2015-11-29





自己紹介

- 佐藤博之
 - Twitter: @hiroysato
 - GitHub: hiroyuki-sato
- 営業・インフラ担当



最近の営業活動

- OS
 - Linux, Windows, OSX
- Infiniband/RoCE
 - RDMA, iSER, SRP
- その他
 - VyOS、 Asterisk、 Sisimai、 PostgreSQL



目次

- 構築したサービスの要件
- Groongaを選んだ理由
- システム構成
- Groongaへデータ登録方法
- 使ってみて

構築したサービスの要件

- 入力ソース
 - 電子メール
- 出力
 - 指定したキーワードの含まれる記事をピックアップ
 - PDF・Excelなどに整形

構築したサービスの要件

✉ 件名: Groonga Meatup
本文: ぐるんがは…

✉ 件名: MySQLで全文検索
本文: Mroongaを使うと

✉ 件名: PostgreSQL
本文: Pgroonga

✉ 件名: 11月29日は
本文: 肉の日



ユーザ



Report

キーワード
Groonga
ぐるんが
PGroonga
Mroonga



全文検索ソフトウェア

日本語全文検索
-> Groonga



なぜGroonga

- 単体で日本語対応してくれそう
 - 実際単体で全角英字や半角英字を正規化してくれた。
 - Groonga G r o o o n g a など
- 須藤さんがメンテナ
 - ActiveLDAP, rcairoなどの実績

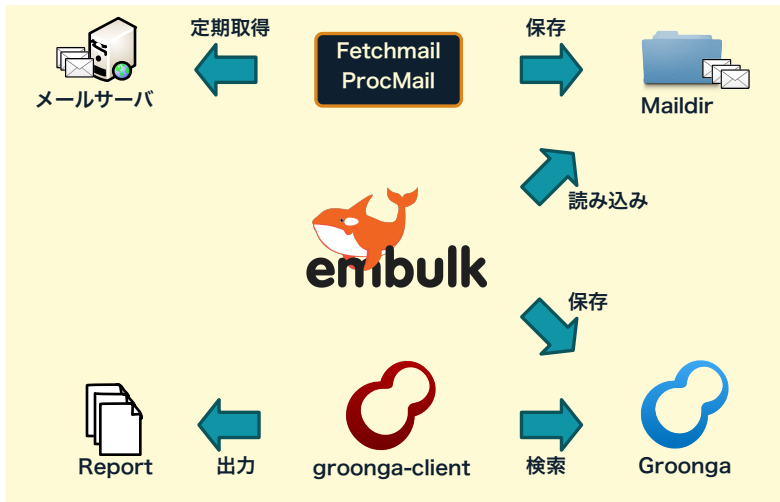


サーバ構成

- 全文検索システム: Groonga
- メールの定期取得
 - fetchmail/procmail
- データ登録: embulk
 - プラグインを自作
- データ出力: groonga-client



構成



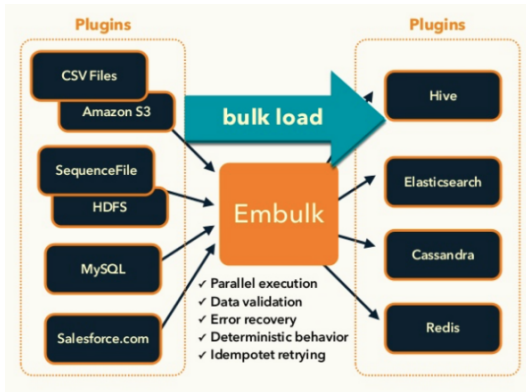


embulk(エンバルク)

- オープンソースバルクローダ
- マルチプラットフォーム
 - OSX, Linux, Windows..
- プラグインアーキテクチャ
 - 約100個のプラグインが利用可能
 - 言語: Java(Scala)とJRuby



embulkの構成



出典：<http://www.slideshare.net/frsyuki/embulk-making-data-integration-works-relaxed>

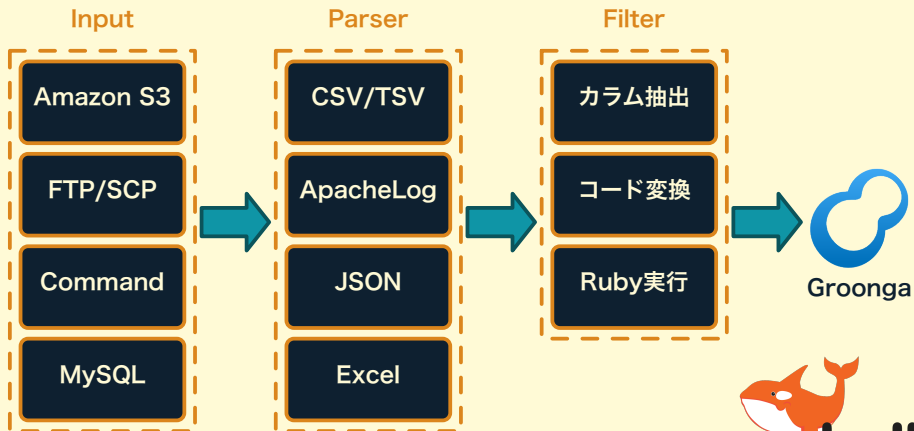


groongaプラグイン

- embulk-output-groonga
 - groongaデータロード用
- ライセンス
 - オープンソース
 - <https://github.com/hiroyuki-sato/embulk-output-groonga>



プラグイン利用例





サンプルデータ

id	title	date	content
1	groonga	2015/11/29	ぐるんが全文検索エンジン
2	pgroonga	2015/11/30	PostgreSQLで全文検索pgroonga
3	mroonga	2015/12/01	MySQLで全文検索Mroonga
4	rroonga	2015/12/02	Rubyで開発rroonga
5	Droonga	2015/12/03	分散groonga Droonga



設定例(入力部)

```
in:
  type: file
  path_prefix: hoge/csv/sample_
  decoders:
  - {type: gzip}
  parser:
    charset: UTF-8
    newline: CRLF
    type: csv
    columns:
    - {name: id, type: long}
    - {name: title, type: string}
    - {name: date, type: timestamp, format: '%Y/%m/%d'}
    - {name: comment, type: string}
```




設定例(出力部)

```
out:  
  type: groonga  
  table: Data # 投入先のテーブル名  
  host: localhost  
  protocol: http  
  key_column: title # キーにするカラム
```



Embulkの情報サイト

- Embulk
 - <http://www.embulk.org>
- Qiita: Fluentdのバッチ版
Embulk(エンバルク)のまとめ
 - <http://qiita.com/hiroysato/items/397f36c4838a0a93e352>



導入結果

なんか遅い



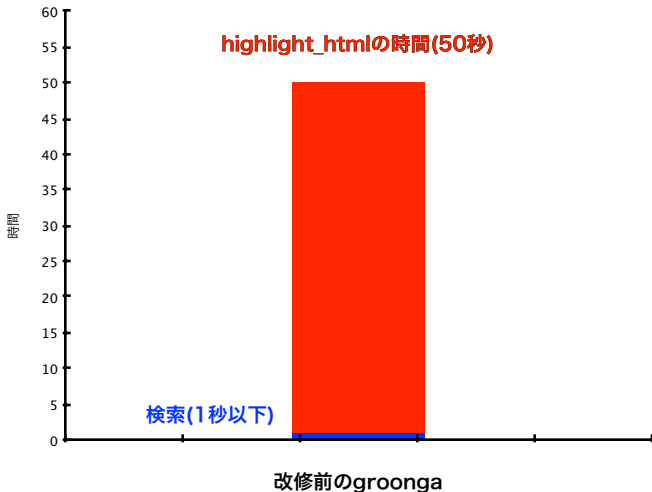
疑似コード

```
select \  
  --table Data \  
  --match_columns article \  
  --output_columns "_key,highlight_html(article),line_no" \  
  --query "( Groonga OR ぐるんが ) OR (line_no:>1 + line_no:<500)" \  
  --command_version 2 \  
  --limit -1
```

Groongaなどのキーワードは40個
ぐらい



遅い原因





遅い原因

highlight_html
がとっても遅い



highlight_html

- **ぐるんが**は全文検索エンジンです。
- ぐるなびさんでも活用されている**Groonga**
- 毎年11月29日に開催される**Groonga**の会



不具合報告

須藤さん助けて



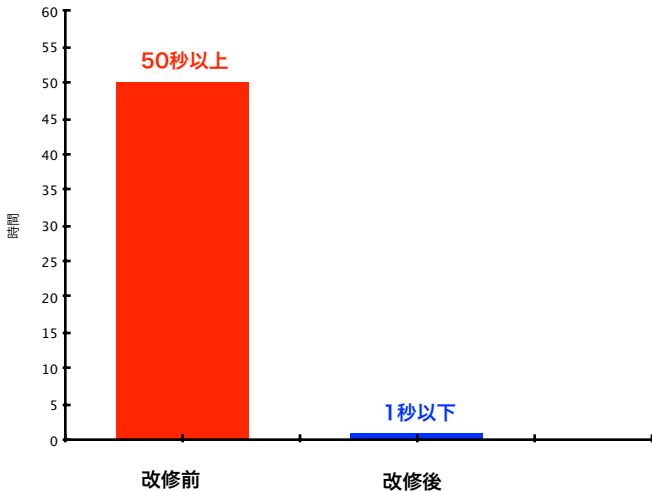
不具合報告

The screenshot shows a GitHub repository page for 'hiroyuki-sato / groonga-highlight_test'. The repository has 16 commits, 1 branch, 0 releases, and 1 contributor. The commit history table is as follows:

Commit	Message	Time
hiroyuki-sato Add query test3		Latest commit bd51c51 on Sep 3
analze	Add query test3	3 months ago
data	Add new data	3 months ago
database	first import	3 months ago
log	first import	3 months ago
log_archives	Add query test3	3 months ago
.gignore	Add gignore	3 months ago
Makefile	first import	3 months ago
README.md	Update README	3 months ago



改修結果(5.0.8~)





Thanks

ありがとうございました。