

Groonga and Apache Arrow

Abe Tomoaki
ClearCode Inc.

**Groongaは
Apache Arrow
フォーマットで入
出力できるよ！**

Groongaとは？

- ✓ オープンソースのカラムストア
機能付き全文検索エンジン
- ✓ 即時更新
- ✓ 参照ロックフリー
- ✓ 集計もできる

例で紹介Groonga: 検索

```
// SQLで書くこういう感じ
// select _key, introduction
// from Users
// where introduction &@ 'engine'
query = new URLSearchParams({
  table: 'Users',
  output_columns: '_key,introduction',
  match_columns: 'introduction',
  query: 'engine', // 「engine」で全文検索
  // output_type: 'apache-arrow', このパラメータを付与するとApache Arrowフォーマットになる
  command_version: 3
});
response = await fetch(`http://localhost:10041/d/select?${query}`);
(await response.json()).body.drilldowns.department.records
// [
//   [ "Groonga", "Groonga is an open-source fulltext search engine and column store." ],
//   [ "Mroonga", "A MySQL pluggable storage engine based on Groonga." ]
// ]
```

例で紹介Groonga: 集計

```
// SQLで書くところの感じ
// select department, avg(age)
// from Users
// group by department

// drilldownというので集計できる
query = new URLSearchParams({
  table: 'Users',
  output_columns: '_id',
  drilldown: 'department',
  drilldown_calc_types: 'AVG',
  drilldown_calc_target: 'age',
  drilldown_output_columns: '_key,_avg',
  command_version: 3
});
response = await fetch(`http://localhost:10041/d/select?${query}`);
(await response.json()).body.drilldowns.department.records
// [
//   [ "A", 37 ],
//   [ "B", 15.5 ]
// ]
```

事例紹介

活用事例1: Groonga delta

- ✓ MySQL/MariaDBのデータをリアルタイムでGroongaに同期するツール
- ✓ 差分情報をApache Parquet形式で保存
- ✓ Groongaへの取り込みを高速に行える

Groonga delta補足

- ✓ MySQL/MariaDBのbinlog（差分情報）を一度ストレージに書き出す
- ✓ これにより途中で失敗してもやり直しがしやすい
- ✓ この書き出すフォーマットがApache Parquet形式

活用事例2: ログ収集製品

- ✓ 某ログ収集製品でGroongaが利用されている
- ✓ データのやり取りにApache Arrowフォーマットを使うことで高速化できた

まとめ

- ✓ GroongaはApache Arrowフォーマットで入出力できるよ！
- ✓ Groongaは全文検索も集計もできるよ！

宣伝

何かとApache Parquet形式でデータを保存する時代なので、Groongaでデータ分析の前処理とかしてみるといいよ！

参考

✓ Groonga

✓ <https://groonga.org/>

✓ Groonga delta

✓ <https://github.com/groonga/groonga-delta>

✓ <https://www.clear-code.com/blog/2022/5/20/groonga-delta.html>