Red Arrow

Ruby and Apache Arrow

Sutou Kouhei

ClearCode Inc.

RubyKaigi Takeout 2021

2021-09-11

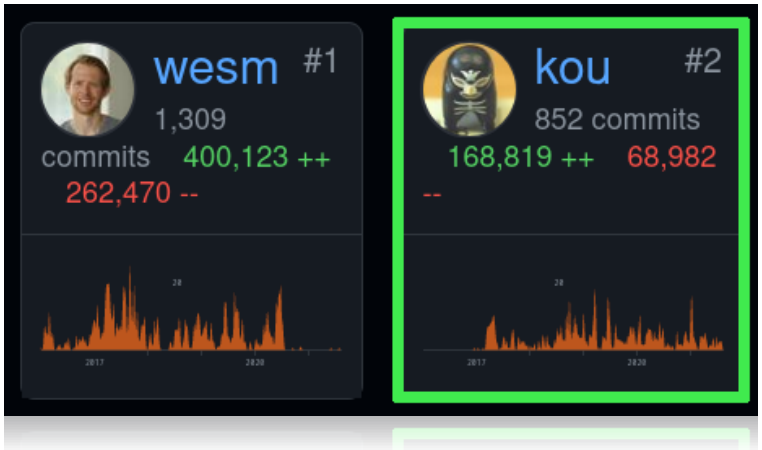
Sutou Kouhei
A president Rubyist

The president of ClearCode Inc.

クリアコードの社長

Sutou Kouhei
An Apache Arrow contributor

- ✓ A member of PMC of Apache Arrow
  PMC: Project Management Committee
  Apache Arrowのプロジェクト管理委員会メンバー
- ✓ #2 commits（コミット数2位）



Sutou Kouhei
The pioneer in Ruby and Arrow

- ✓ The author of Red Arrow
  Red Arrowの作者
- ✓ Red Arrow:
  - ✓ The official Apache Arrow library for Ruby
    公式のRuby用のApache Arrowライブラリー
  - ✓ GObject Introspection based bindings
    GObject Introspectionベースのバインディング
  - ✓ Apache Arrow GLib is developed for Red Arrow
    Red ArrowのためにApache Arrow GLibも開発

GObject Introspection?

A way to implement bindings

バインディングの実装方法の1つ

Ruby bindings 2016

How to create bindings 2016

Kouhei Sutou

ClearCode Inc.

RubyKaigi 2016
2016-09-09

https://rubykaigi.org/2016/presentations/ktou.html

Why do I work on Red Arrow?
なぜRed Arrowの開発をしているか

- ✓ To use Ruby for data processing!
  データ処理でRubyを使いたい！
  - ✓ At least a part of data processing
    データ処理の全部と言わず一部だけでも
- ✓ Results of my 5 years of work:
  私のここ5年の仕事の成果
  - ✓ We can use Ruby for some data processing!
    いくつかのデータ処理でRubyを使える！

Goal of this talk
このトークのゴール

- ✓ You want to use Ruby for some data processing
  いくつかのデータ処理でRubyを使いたくなる
- ✓ You join Red Data Tools project
  Red Data Toolsプロジェクトに参加する

Red Data Tools project?

> *Red Data Tools is a project that provides data processing tools for Ruby*

Red Data ToolsはRuby用のデータ処理ツールを提供するプロジェクト

https://red-data-tools.github.io/

Data processing?

... how?

0. Why do you want?

0. データ処理の目的を明らかにする

- ✓ What problem do you want to resolve?
  どんな問題を解決したい？
- ✓ What data is needed for it?
  そのためにはどんなデータが必要？
- ✓ ...

No Red Arrow support in this area
このあたりにはRed Arrowを使えない

1. Collect data
1. データ収集

- ✓ Where are data?
  データはどこにある？
- ✓ Where are collected data stored?
  集めたデータはどこに保存する？
- ✓ ...

Some Red Arrow supports in this area
このあたりでは少しRed Arrowを使えない

Common dataset
よく使われるデータセット

```
require "datasets"
Datasets::Iris.new
Datasets::PostalCodeJapan.new
Datasets::Wikipedia.new
```

Red Datasets
https://github.com/red-data-tools/red-datasets

Output: Local file
出力先：ローカルファイル

```
require "datasets-arrow"
dataset = Datasets::PostalCodeJapan.new
dataset.to_arrow.save("codes.csv")
dataset.to_arrow.save("codes.arrow")
```

Red Datasets Arrow
https://github.com/red-data-tools/red-datasets-arrow

#save

- General serialize API for table data
  テーブルデータ用の汎用シリアライズAPI
  - Serialize as the specified format
    指定したフォーマットにシリアライズ
  - If you use Red Arrow object for in-memory table data, you can serialize to many formats! Cool!
    メモリー上のテーブルデータをRed Arrowオブジェクトにするといろんなフォーマットにシリアライズできる！かっこいい！
- Extensible!
  拡張可能！

#save: Implementation

```
module Arrow
  class Table
    def save(output)
      saver = TableSaver.new(self, output)
      saver.save
    end
  end
end
```

#save: Implementation

```
class Arrow::TableSaver
  def save
    format = detect_format(@output)
    __send__("save_as_#{format}")
  end
  def save_as_csv
  end
end
```

#save: Extend by Red Parquet

```
module Parquet::ArrowTableSavable
  def save_as_parquet
  end
  Arrow::TableSaver.include(self)
end
```

Red Parquet is a subproject of Red Arrow
Red ParquetはRed Arrowのサブプロジェクト

#save: Extended

```
require "datasets-arrow"
require "parquet"
dataset = Datasets::PostalCodeJapan.new
dataset.to_arrow.save("codes.parquet")
```

Output: Online storage: Fluentd
出力先：オンラインストレージ：Fluentd

- ✓ fluent-plugin-s3-arrow:
  - ✓ Collect data by Fluentd
    Fluentdでデータ収集
  - ✓ Format data as Apache Parquet by **Red Arrow**
    Red ArrowでApache Parquet形式にデータを変換
  - ✓ Store data to Amazon S3 by fluent-plugin-s3
    fluent-plugin-s3でAmazon S3にデータを保存
  - ✓ By @kanga33 at Speee/Red Data Tools
    Speee/Red Data Toolsの香川さんが開発

https://github.com/red-data-tools/fluent-plugin-s3-arrow/

Output: Online storage: Red Arrow

出力先：オンラインストレージ：Red Arrow

```
require "datasets-arrow"
require "arrow-dataset"
dataset = Datasets::PostalCodeJapan.new
url = URL("s3://mybucket/codes.parquet")
dataset.to_arrow.save(url)
```

Implementing...

実装中。。。

#save: Implementing...

```
class Arrow::TableSaver
  def save
    if @output.is_a?(URI)
      __send__("save_to_uri")
    else
      __send__("save_to_file")
    end
  end
end
```

Collect data w/ Red Arrow: Wrap up

Red Arrowでデータ収集：まとめ

- ✓ Usable as serializer for common formats
  よくあるフォーマットにシリアライズするツールとして使える
- ✓ Usable as writer to common locations
  in the near future...
  近いうちによくある出力先に書き出すツールとして使える

2. Read data
2. データ読み込み

- ✓ What format is used?
  どんなフォーマットで保存されている？
- ✓ Where are collected data?
  収集したデータはどこ？
- ✓ How large is collected data?
  データはどれかで大きい？

Format
フォーマット

```
require "arrow"
table = Arrow::Table.load("data.csv")
table = Arrow::Table.load("data.json")
table = Arrow::Table.load("data.arrow")
table = Arrow::Table.load("data.orc")
```

.load

- ✓ General deserialize API for table data
  テーブルデータ用の汎用デシリアライズAPI
  - ✓ Deserialize common formats
    よく使われているフォーマットからデシリアライズ
- ✓ Extensible!
  拡張可能！

.load: Implementation

```
module Arrow
  def Table.load(input)
    loader = TableLoader.new(self, input)
    loader.load
  end
end
```

.load: Implementation

```
class Arrow::TableLoader
  def load
    format = detect_format(@output)
    __send__("load_as_#{format}")
  end
  def load_as_csv
  end
end
```

.load: Extend by Red Parquet

```
module Parquet::ArrowTableLoadable
  def load_as_parquet
  end
  Arrow::TableLoader.include(self)
end
```

Red Parquet is a subproject of Red Arrow
Red ParquetはRed Arrowのサブプロジェクト

.load: Extended

```ruby
require "parquet"
table = Arrow::Table.load("data.parquet")
```

.load: More extensible

```
class Arrow::TableLoader
  def load
    if @output.is_a?(URI)
      __send__("load_from_uri")
    else
      __send__("load_from_file")
    end
  end
end
```

.load: Extend by Red Arrow Dataset

```
module ArrowDataset::ArrowTableLoadable
  def load_from_uri
  end
  Arrow::TableLoader.include(self)
end
```

Red Arrow Dataset is a subproject of Red Arrow
Red Arrow DatasetはRed Arrowのサブプロジェクト

Location: Online storage
場所：オンラインストレージ

```
require "arrow-dataset"
url = URI("s3://bucket/path...")
table = Arrow::Table.load(url)
```

Location: RDBMS

場所：RDBMS

```
require "arrow-activerecord"
User.all.to_arrow
```

Red Arrow Active Record
https://github.com/red-data-tools/red-arrow-activerecord

Location: Network
場所：ネットワーク

```
require "arrow-flight"
client = ArrowFlight::Client.new(url)
info = client.list_flights[0]
reader = client.do_get(info.endpoints[0].ticket)
table = reader.read_all
```

Introducing Apache Arrow Flight: A Framework for Fast Data Transport
https://arrow.apache.org/blog/2019/10/13/introducing-arrow-flight/

Large data
大規模データ

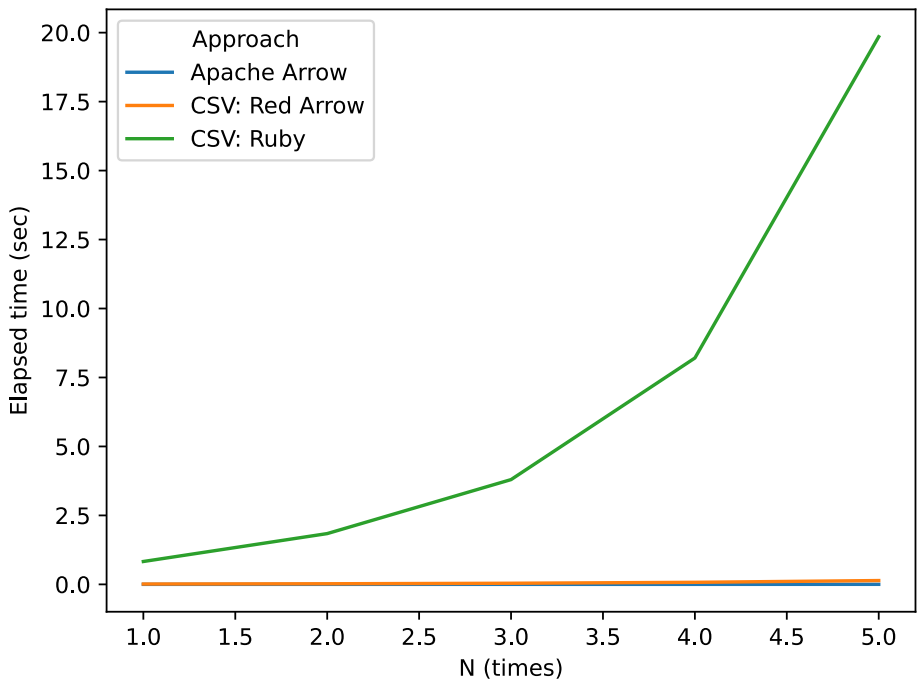
- Apache Arrow format
  - Designed for large data
    大規模データ用に設計されている
- For large data
  大規模データ用に必要なもの
  - Fast load
    高速にロードできること
  - ...

Fast load: Benchmark
高速ロード：ベンチマーク

```ruby
require "datasets-arrow"
dataset = Datasets::PostalCodeJapan.new
table = dataset.to_arrow # 124271 records
n = 5
n.times do |i|
  table.save("codes.#{i}.csv")
  table.save("codes.#{i}.arrow")
  CSV.read("codes.#{i}.csv")
  Arrow::Table.load("codes.#{i}.csv")
  Arrow::Table.load("codes.#{i}.arrow")
  table = table.concatenate([table])
end
```

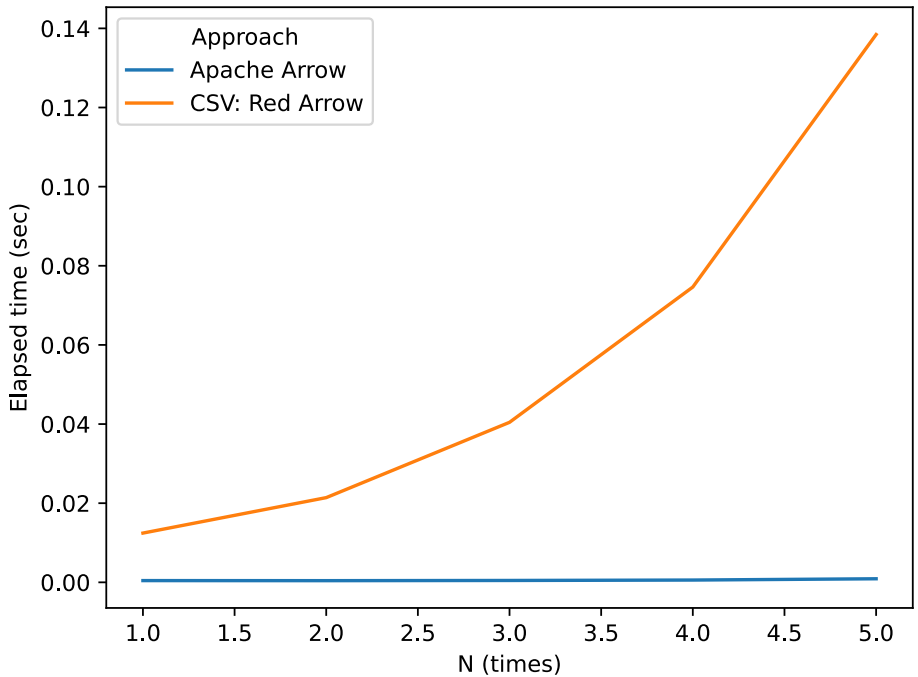
Fast load: Benchmark: All

高速ロード：ベンチマーク：すべて



Fast load: Benchmark: Red Arrow

高速ロード：ベンチマーク：Red Arrow



How to implement fast load
高速ロードの実装方法

ClearCode

Apache Arrowフォーマットは なぜ速いのか

須藤功平
株式会社クリアコード

db tech showcase ONLINE 2020
2020-12-08

https://slide.rabbit-shocker.org/authors/kou/db-tech-showcase-online-2020/

Read data with Red Arrow: Wrap up

Red Arrowでデータ読み込み：まとめ

- ✓ Easy to read common formats
  よくあるフォーマットのデータを簡単に読める
- ✓ Easy to read from common locations
  よくある場所にあるデータを簡単に読める
- ✓ Large data ready
  大規模データも扱える

3. Explore data
3. データ探索

- ✓ Preprocess data（データを前処理）
  - ✓ Filter out needless data（不要なデータを除去）
  - ✓ ...
- ✓ Summarize data and visualize them
  （データを要約して可視化）
- ✓ ...

Red Arrow can be used for some operations
いくつかの操作でRed Arrowを使える

Filter: Red Arrow
絞り込み：Red Arrow

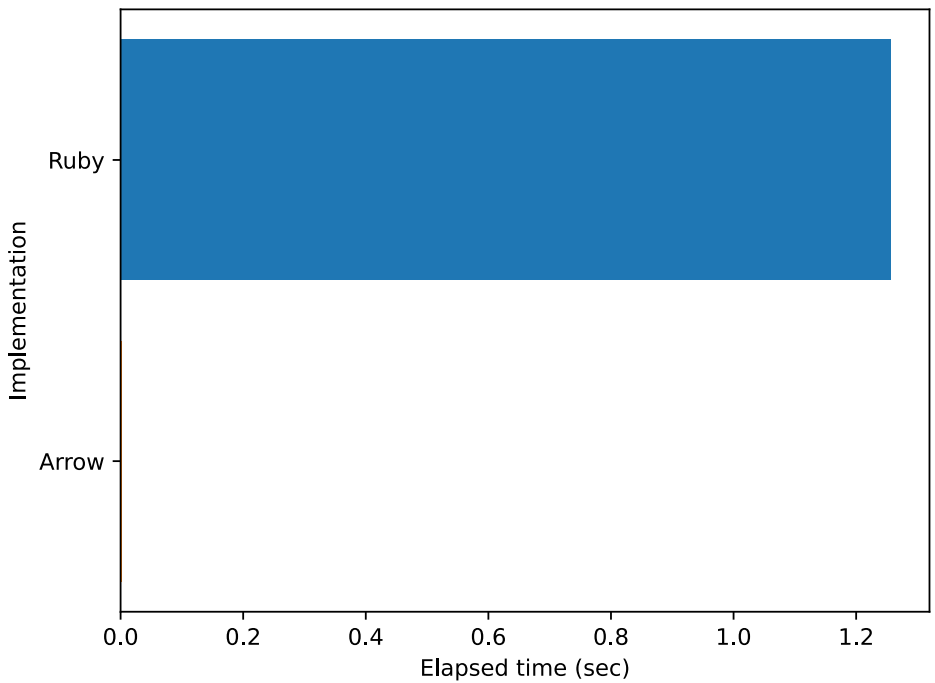
```
table = Datasets::PostalCodeJapan.new.to_arrow
table.n_rows # 124271
filtered_table = table.slice do |slicer|
  slicer.prefecture == "東京都" # Tokyo
end
filtered_table.n_rows # 3887
```

Filter: Performance
絞り込み：性能

```
dataset = Datasets::PostalCodeJapan.new
arrow_dataset = dataset.to_arrow
dataset.find_all do |row|
  row.prefecture == "東京都" # Tokyo
end # 1.256s
arrow_dataset.slice do |slicer|
  slicer.prefecture == "東京都" # Tokyo
end # 0.001s
```

Filter: Performance

絞り込み：性能



Apache Arrow data: Interchangeable

Apache Arrow data：交換可能

- With low cost thanks to fast load
  高速ロードできるので低コスト
- Apache Arrow data ready systems are increasing
  Apache Arrowデータを扱えるシステムは増加中
  - e.g. DuckDB: in-process SQL OLAP DBMS
    (SQLite like DBMS for OLAP)
    OLAP: OnLine Analytical Processing
    例：DuckDB：同一プロセス内で動くデータ分析用SQL DB管理システム

Filter: DuckDB

絞り込み：DuckDB

```
require "arrow-duckdb"
codes = Datasets::PostalCodeJapan.new.to_arrow
db = DuckDB::Database.open
c = db.connect
c.register("codes", codes) do # Use codes without copy
  c.query("SELECT * FROM codes WHERE prefecture = ?",
          "東京都", # Tokyo
          output: :arrow) # Output as Apache Arrow data
   .to_table.n_rows # 3887
end
```

Summarize: Group + aggregation

要約：グループ化して集計

```
iris = Datasets::Iris.new.to_arrow
iris.group(:label).count(:sepal_length)
#       count(sepal_length)     label
# 0                        50     Iris-setosa
# 1                        50     Iris-versicolor
# 2                        50     Iris-virginica
```

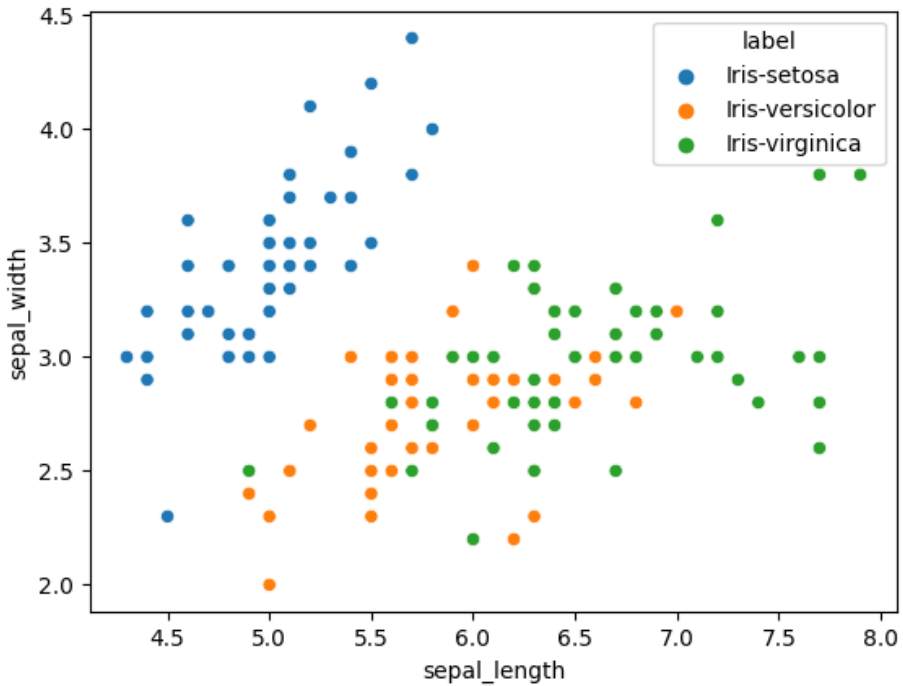
Visualize: Charty

可視化：Charty

```
require "charty"
Charty.backends.use("pyplot")
Charty.scatter_plot(data: iris,
                    x: :sepal_length,
                    y: :sepal_width,
                    color: :label)
      .save("iris.png")
```

Visualize: Charty: Result

可視化：Charty：結果



4. Use insight
4. 知見を活用

- ✓ Write report
  （レポートにまとめたり）
- ✓ Build a model
  （モデルを作ったり）
- ✓ ...

No Red Arrow support in this area for now

Can be used for passing data to other tools like DuckDB and Charty
今のところこのあたりにはRed Arrowを使えない
DuckDBやChartyにデータを渡すように他のツールにデータを渡すためには使える

Data processing and Red Arrow
Red Arrowでデータ処理

- ✓ Red Arrow helps us in some areas
  いくつかの領域ではRed Arrowを使える
  - ✓ Collect, read and explore data
    データを収集して読み込んで探索するとか
- ✓ Some tools can integrate with Red Arrow
  いくつかのツールはRed Arrowと連携できる
  - ✓ Fluentd, DuckDB, Charty, ...

Red Arrow and Ruby 3.0

- ✓ MemoryView support
- ✓ Ractor support

MemoryView

> *MemoryView provides the features to share multidimensional homogeneous arrays of fixed-size element on memory among extension libraries.*

MemoryViewは多次元数値配列（数値はすべて同じ型）を共有する機能を提供します。

https://docs.ruby-lang.org/en/master/doc/memory_view_md.html

https://tech.speee.jp/entry/2020/12/24/093131 (Japanese)

Numeric arrays in Red Arrow

Red Arrow内の数値配列

- ✓ Arrow::NumericArray family
  - ✓ 1-dimensional numeric array
    1次元数値配列
- ✓ Arrow::Tensor
  - ✓ Multidimensional homogeneous numeric arrays
    多次元数値配列

MemoryView: Red Arrow

- ✓ Arrow::NumericArray family
  - ✓ Export as MemoryView: Support
    MemoryViewとしてエクスポート：対応済み
  - ✓ Import from MemoryView: Not yet
    MemoryViewをインポート：未対応
- ✓ Arrow::Tensor
  - ✓ Export/Import: Not yet
    エクスポート・インポート：未対応

Join Red Data Tools to work on this!
対応を進めたい人はRed Data Toolsに来てね！

MemoryView: C++

- ✓ Some problems are found by this work
  Red Arrowの対応作業でいくつかの問題が見つかった
  - ✓ Can't use private as member name
    メンバー名にprivateを使えない
  - ✓ Can't assign to const variable with cast
    キャストしてもconst変数に代入できない
- ✓ Ruby 3.1 will fix them
  Ruby 3.1では直っているはず

Ractor

> *Ractor is designed to provide a parallel execution feature of Ruby without thread-safety concerns.*

Ractorはスレッドセーフかどうかを気にせずに並列実行するための機能です。

https://docs.ruby-lang.org/en/master/doc/ractor_md.html

https://techlife.cookpad.com/entry/2020/12/26/131858 (Japanese)

Red Arrow and concurrency

Red Arrowと並列性

- ✓ Red Arrow data are immutable
  Red Arrowデータは変更不可
- ✓ Ractor can share frozen objects
  RactorはfrozenなオブジェクトをShareable可能

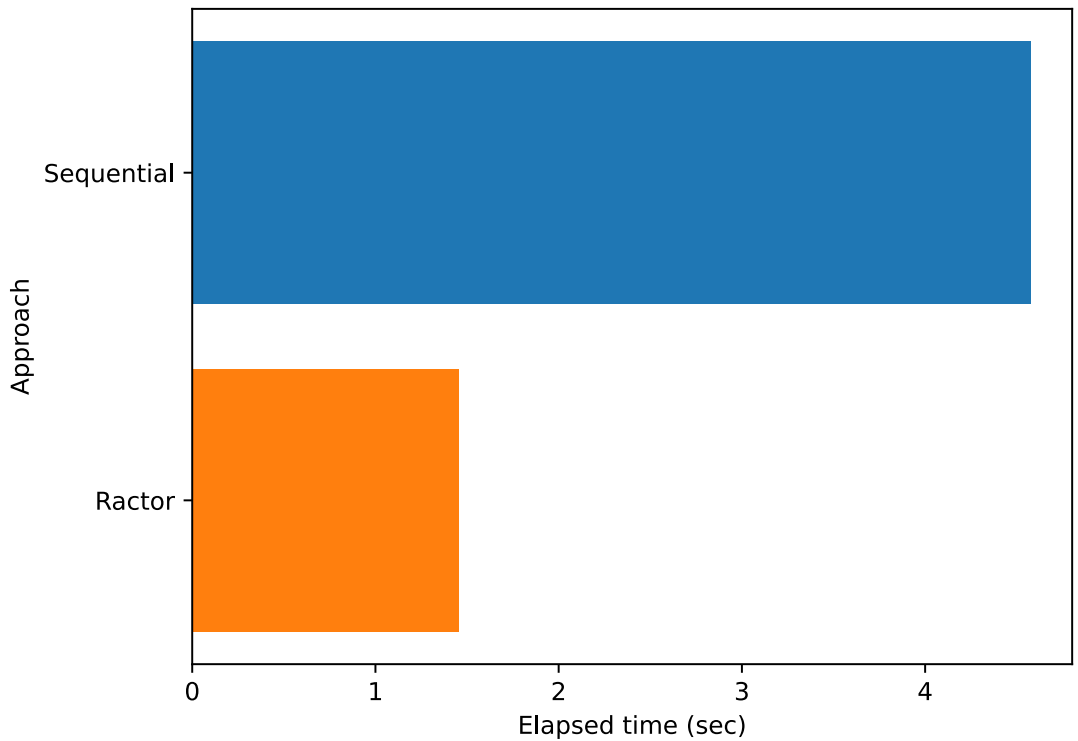
Ractor: Red Arrow

```ruby
require "datasets-arrow"
table = Datasets::PostalCodeJapan.new.to_arrow
Ractor.make_shareable(table)
Ractor.new(table) do |t|
  t.slice do |slicer|
    slicer.prefecture == "東京都" # Tokyo
  end
end
```

Ractor: Red Arrow: Benchmark

```
n_ractors = 4
n_jobs_per_ractor = 1000
n_jobs = n_ractors * n_jobs_per_ractor
n_jobs.times do
  table.slice {|s| s.prefecture == "東京都"}
end
n_ractors.times.collect do
  Ractor.new(table, n_jobs_per_ractor) do |t, n|
    n.times {t.slice {|s| s.prefecture == "東京都"}}
  end
end.each(&:take)
```

Ractor: Red Arrow: Benchmark



Wrap up
まとめ

- ✓ Ruby can be used in some data processing work
  いくつかのデータ処理作業にRubyを使える
  - ✓ Red Arrow helps you!
    Red Arrowが有用なケースがあるはず！
- ✓ Ruby 3.0 has useful features for data processing work
  Ruby 3.0にはデータ処理作業に有用な機能があるよ
  - ✓ Red Arrow starts supporting them
    Red Arrowはそれらのサポートを進めている

Goal of this talk
このトークのゴール

- ✓ You want to use Ruby for some data processing
  いくつかのデータ処理でRubyを使いたくなる
- ✓ You join Red Data Tools project
  あなたがRed Data Toolsプロジェクトに参加する

Feature work
今後の仕事

- ✓ Implement DataFusion bindings by adding C API to DataFusion
  DataFusionにC APIを追加してバインディングを実装
  - ✓ DataFusion: Apache Arrow native query execution framework written in Rust
    https://github.com/apache/arrow-datafusion/
    DataFusion：Rust実装のApache Arrowベースのクエリー実行フレームワーク
- ✓ Add Active Record like API to Red Arrow
  Red ArrowにActive Record風のAPIを追加
- ✓ Improve MemoryView/Ractor support
  MemoryView/Ractorサポートを進める

Red Data Tools

Join us!

https://red-data-tools.github.io/
https://gitter.im/red-data-tools/en
https://red-data-tools.github.io/ja/
https://gitter.im/red-data-tools/ja

OSS Gate on-boarding
OSS Gateオンボーディング

- ✓ Supports accepting newcomers by OSS projects such as Ruby & Red Arrow
  RubyやRed ArrowといったOSSプロジェクトが新人を受け入れることを支援
- ✓ Contact me! 興味がある人は私に教えて！
  - ✓ OSS project members who want to accept newcomers
    新人を受け入れたいOSSプロジェクトのメンバー
  - ✓ Companies which want to support OSS Gate on-boarding
    OSS Gateオンボーディングを支援したい会社

https://oss-gate.github.io/on-boarding/

ClearCode Inc.

- Recruitment: Developer to work on Red Arrow related business
  採用情報：Red Arrow関連のビジネスをする開発者
  - https://www.clear-code.com/recruitment/
- Business: Apache Arrow/Red Arrow related technical support/consulting:
  仕事：Apache Arrow/Red Arrow関連の技術サポート・コンサルティング
  - https://www.clear-code.com/contact/