Why Apache Arrow is important for Ruby

Sutou Kouhei

ClearCode Inc.

The Data Thread

2022-06-23

Me

- ✓ Name: Sutou Kouhei
  (Family Given)
- ✓ ID: kou (call me kou)
  (ktou or kous when I can't use kou)
- ✓ Ruby committer since 2004
- ✓ This year's Apache Arrow PMC chair

My profile picture is my "Shocker combatman" figure on my Happy Hacking Keyboard

Why I work on Apache Arrow

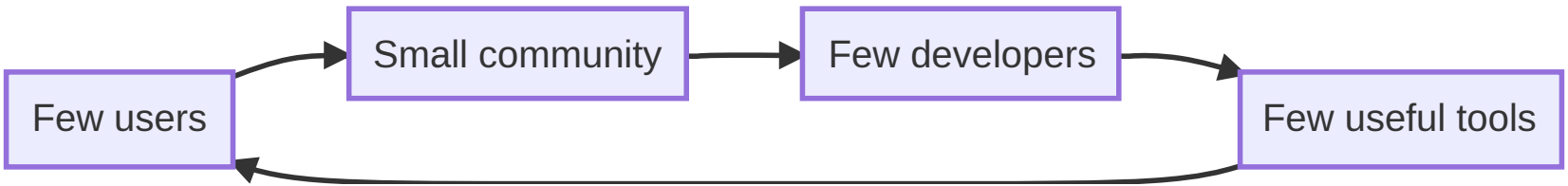
For Ruby!
(I love Ruby!)

Ruby

- ✓ Widely used for Web application
  (I rarely write Web app)
  - ✓ Ruby on Rails is an useful Web app framework
  - ✓ e.g.: GitHub, GitLab, Shopify, Discourse, ...
- ✓ Not widely used for data processing
  - ✓ Even though Ruby is a general purpose programming language...

Ruby and data processing

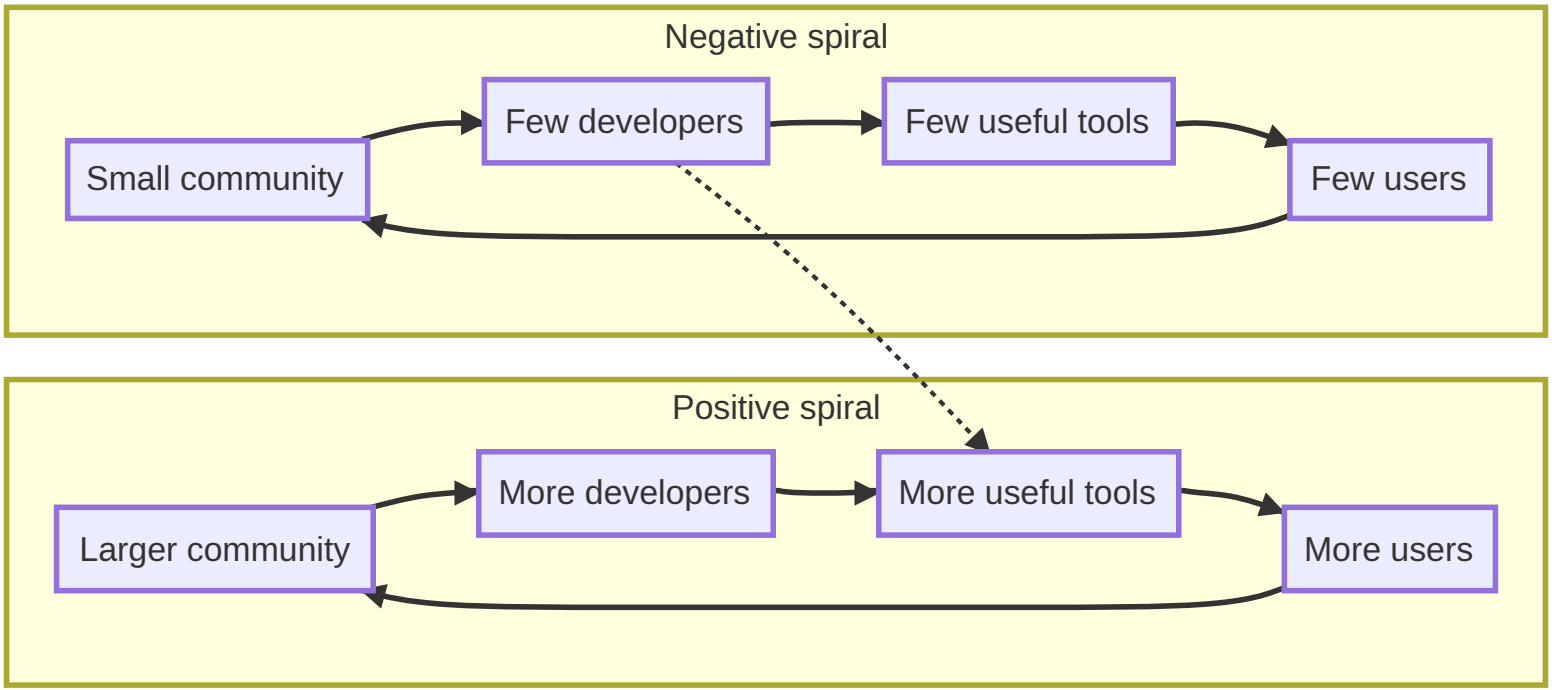
Negative spiral



How to break the negative spiral?

- ✓ Few users: Expand useful tools?
- ✓ Small community: Increase # of users?
- ✓ Few developers: Expand community?
- ✓ Few useful tools: Increase # of developers?

Expand useful tools with few developers



But how?

Apache Arrow

Apache Arrow

- Cross-language dev platform
  - Ruby community doesn't need to dev everything
  - We can share common implementations
- Apache Arrow and Ruby
  - I've donated the Ruby bindings for C++ in 2017
  - Ruby bindings: Red Arrow
  - Many features are already bound:
    Parquet, Dataset, Gandiva, Flight, ...

Red Data Tools

I started a new project in 2017:

> *Red Data Tools is a project that provides data processing tools for Ruby.*
>
> *[cited from `https://red-data-tools.github.io/']*

Red Data Tools: Policy 1

Collaborate across the Ruby community

We collaborate with the Ruby community and other communities. For example, we use Apache Arrow, shared with many languages, and join in development of Apache Arrow to share benefits.

[cited from `https://red-data-tools.github.io/']

What fields I work on

- ✓ Not only Ruby related features
  - ✓ To be a good Apache Arrow community member
- ✓ Community support
  - ✓ Answer questions from users
  - ✓ Review pull requests

What features I work on

- ✓ Ruby related
  - ✓ C++ impl., C GLib bindings, Linux packages, Homebrew, MSYS2, Release, CI, ...
- ✓ Not Ruby related
  - ✓ wheel, jar, MATLAB bindings, Julia impl., ...

What fields Red Data Tools members work on

- ✓ C GLib bindings
- ✓ Red Arrow
- ✓ Tensor
- ✓ Big endian
- ✓ C++ compute functions

What skills I have

not used for Apache Arrow yet

Develop MySQL/PostgreSQL plugin

- ✓ I'm a developer of Mroonga/PGroonga
  - ✓ Mroonga: A MySQL plugin for full text search
    (mú lúngá)
  - ✓ PGroonga: A PG plugin for full text search
    (pí:zí: lúngá)
- ✓ Use case: Impl. Flight SQL adapter?

and more...

Apache Arrow and Ruby community

- ✓ Ruby community uses Arrow's work
- ✓ Ruby community joins in Arrow dev

What feature is useful for Ruby?

Fast data interchange

Fast data interchange

- ✓ It's still difficult to use Ruby for full data processing
  - ✓ Because Apache Arrow doesn't solve everything
- ✓ Increase usage of Ruby step by step
  - ✓ Because Ruby can integrate with other languages by Apache Arrow's fast data interchange feature

Integration examples

- DuckDB:
  Arrow ready in-process SQL OLAP DBMS
  - https://github.com/red-data-tools/red-arrow-duckdb
- DataFusion:
  Arrow native SQL query engine
  - WIP: Export C API #1113
    https://github.com/apache/arrow-datafusion/issues/1113

What feature is useful for Ruby?

Web app related features

Because many Ruby users develop Web apps with Ruby on Rails

What features are useful for Web app

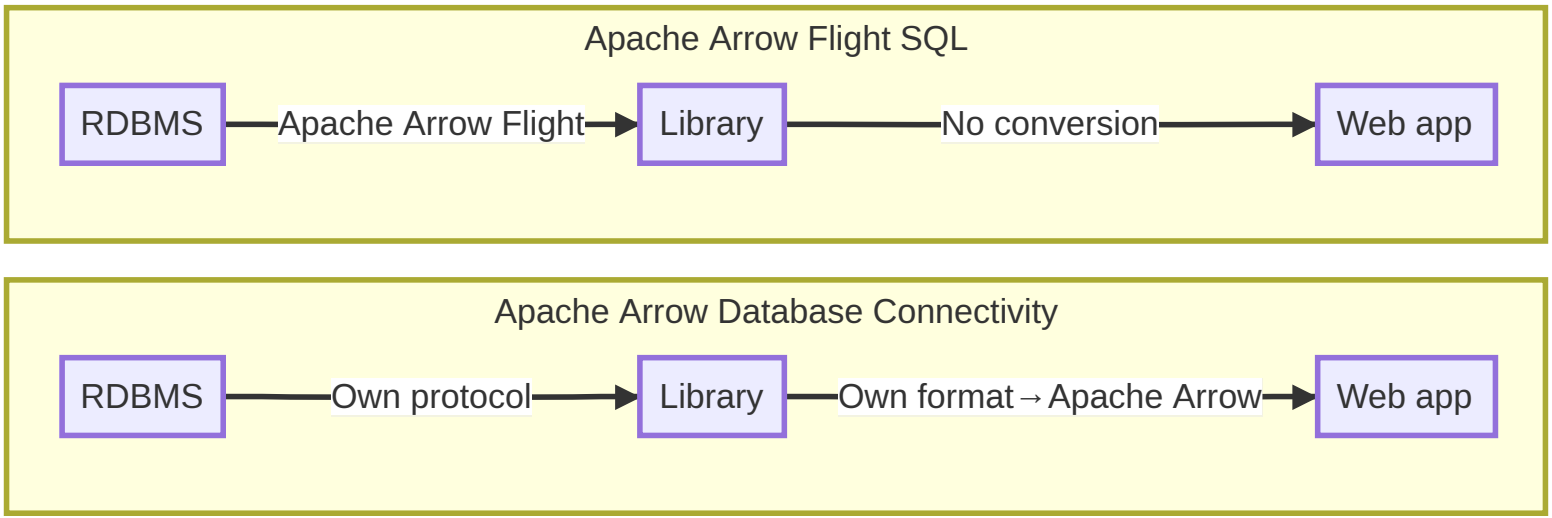
- ✓ Visualization related features
  - ✓ For dashboard
- ✓ Fast data interchange with RDBMS
  - ✓ Web app may have batch jobs to process large data in RDBMS
  - ✓ See also: mrkn's talk on RubyKaigi 2019
    (mrkn is an Apache Arrow committer from Red Data Tools)
    https://speakerdeck.com/mrkn/reducing-activerecord-memory-consumption-using-apache-arrow

Fast data interchange with RDBMS

- ✓ Apache Arrow Flight SQL
- ✓ Apache Arrow Database Connectivity: ADBC

https://docs.google.com/document/d/1t7NrC7O3yxL_0ffATmjzZs2xcj1owdUsIF2WKL_Zw1U/

Fast data interchange with RDBMS



Apache Arrow data⇄Ruby objects

- ✓ Red Arrow has fast converter
  - ✓ Implemented in C++
- ✓ Faster than RDBMS's own format data⇄Ruby objects
  - ✓ Both of Flight SQL and ADBC will improve performance

Wrap up

- ✓ Ruby community joins in Arrow dev
  - ✓ To use Ruby for data processing
- ✓ Ruby community is interested in:
  - ✓ Integration with other data processing systems
  - ✓ RDBMS related improvements

Topics I didn't talk today

- GObject Introspection (GI)
  - Ruby bindings are generated at run-time not compile-time
  - How does GI work for it?
- Linux packaging
  - How to build deb/rpm for Debian/Ubuntu/CentOS/AlmaLinux/Amazon Linux on x86_64 and arm64?

Acknowledgment

- Voltron Data
  - Most of my Apache Arrow related work is being done with financial support from Voltron Data since 2022-04
- Yukiko Yoshimoto at ClearCode
  - Add English subtitle to this video