

Apache Arrow

須藤功平

クリアコード

データ分析用次世代データフォーマット

Apache Arrow勉強会

2017-05-28

ハッシュタグ

#osaka_arrow

今日はいろんなURLを参照するのでそれらを共有したい

流れ

1. Apache Arrowの概要を知る
2. Apache Arrowの詳細を知る
3. Apache Arrow関連の開発に参加する方法を知る

概要

DataScience.rbワークショップ
の資料で紹介

- ✓ RubyもApache Arrowで
データ処理言語の仲間入り
[https://slide.rabbit-shocker.org/
authors/kou/data-science-rb/](https://slide.rabbit-shocker.org/authors/kou/data-science-rb/)

詳細

- ✓ 最新情報はWes McKinneyさんのスライドを見るのがよい
 - ✓ <https://www.slideshare.net/wesm/>
- ✓ 例：
 - ✓ <https://www.slideshare.net/wesm/memory-interoperability-in-analytics-and-machine-learning>
 - ✓ <https://www.slideshare.net/wesm/nextgeneration-python-big-data-tools-powered-by-apache-arrow>

紹介

- ✓ https://www.slideshare.net/MapR_Japan/apache-arrow-value-vectors-tokyo-apache-drill-meetup-20160322
- ✓ <https://www.slideshare.net/wesm/nextgeneration-python-big-data-tools-powered-by-apache-arrow>
- ✓ <https://www.slideshare.net/wesm/memory-interopability-in-analytics-and-machine-learning>

開発に参加

- ✓ Apache Arrowの旨味がでる状態
 - ✓ みんながApache Arrowを使う
- ✓ 早く↑の状態にするには
 - ✓ Apache Arrow関連の開発に参加！
待っていることもできるけど一緒にやろうよ！

Apache Arrowの開発に参加

- ✓ JIRA: <https://issues.apache.org/jira/browse/ARROW/>
 - ✓ コミットはすべてチケットに紐づく
 - ✓ こういうのやりたいねー！もチケットになる
- ✓ メーリングリスト: dev@arrow.apache.org
dev-subscribe@arrow.apache.orgにメールを送ればOK
 - ✓ 基本的にここでディスカッション
 - ✓ JIRAの新規チケットも流れる

Apache Arrowの開発に参加

- ✓ バグレポート
 - ✓ JIRAにチケット作成
- ✓ バグ修正・機能追加
 - ✓ JIRAにチケット作成→GitHubでPR
Pull Requestタイトルにルールあり（後述）
- ✓ 相談
 - ✓ メーリングリスト

PRのタイトル

フォーマット :

ARROW-XXX: [YYY] ...

例 :

ARROW-897: [GLib] Extract ...

ARROW-XXX: JIRAのissue ID

[YYY]: モジュール名

モジュール

- ✓ Java: Java実装
- ✓ C++: C++実装
- ✓ GLib: C++実装のCラッパー
(各種言語バインディング向け)
 - ✓ GLibを使用
- ✓ JS: JavaScript実装
 - ✓ TypeScriptを使用

WANTED: モジュール

↓は未着手なはずなので
ここから開発に参加もあり

- ✓ R: C++実装のR_{cpp}ラッパー
- ✓ Go: Go実装
 - ✓ すでにGLib経由で使えるけど
ネイティブ実装の方がいいかも?
- ✓ Rust: Rust実装

Apache Arrow関連の開発

- ✓ 大量のデータ交換が必要な
プロダクトをArrowに対応させる
- ✓ 例：Apache Spark
(PySparkはすでに進んでいる：SPARK-13534)
- ✓ 参考：対応済みプロダクト
 - ✓ Groonga: <http://groonga.org/>
 - ✓ Ray: <https://github.com/ray-project/ray>

Red Data Tools

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

- ✓ Ruby用データ分析ツールを揃えよう！プロジェクト
 - ✓ Apache Arrowベース
- ✓ ただし！できるだけRuby以外でも使えるようにしたい！

Ruby以外でも使える？

- ✓ GLibバインディングとして開発
(Ruby専用バインディングとして開発しない)
- ✓ Luaとかでも使えるようになる
- ✓ 例：parquet-glib
<https://github.com/red-data-tools/parquet-glib>
- ✓ 例：xtensor-glib
<https://github.com/red-data-tools/xtensor-glib>

Ruby以外でも使える？

- ✓ データも似たような感じで
- ✓ どうすればいろんな言語から
使いやすくなるかは要検討

開発に参加しよう！

- ✓ Apache Arrow
 - ✓ dev@arrow.apache.org
- ✓ Red Data Tools
 - ✓ <https://gitter.im/red-data-tools>
- ✓ Arrowが嬉しそうなプロダクト