# Spark 2.0: What's Next

Reynold Xin @rxin

Spark Conference Japan

Feb 8, 2016

Please put up your hand

if you know what Spark is?

Put up your hand

if you think your significant other know what Spark is?

(girlfriend, boyfriend, wife, husband, …)

# This Talk

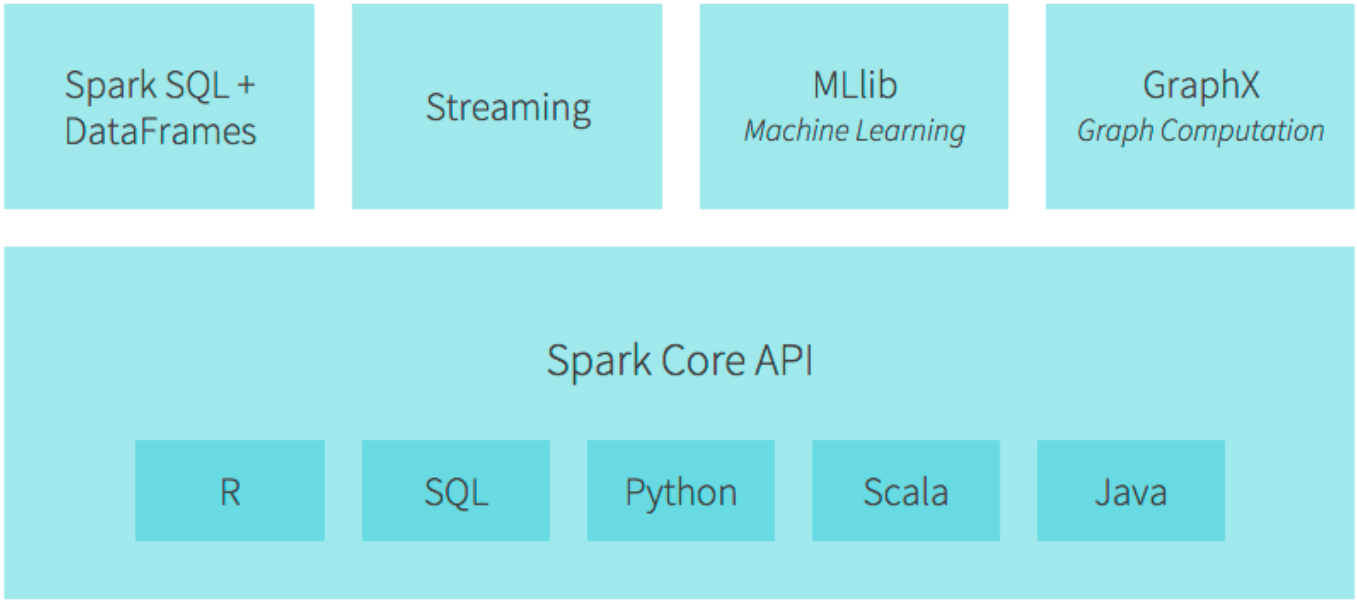
What is Spark?

How are people using it?

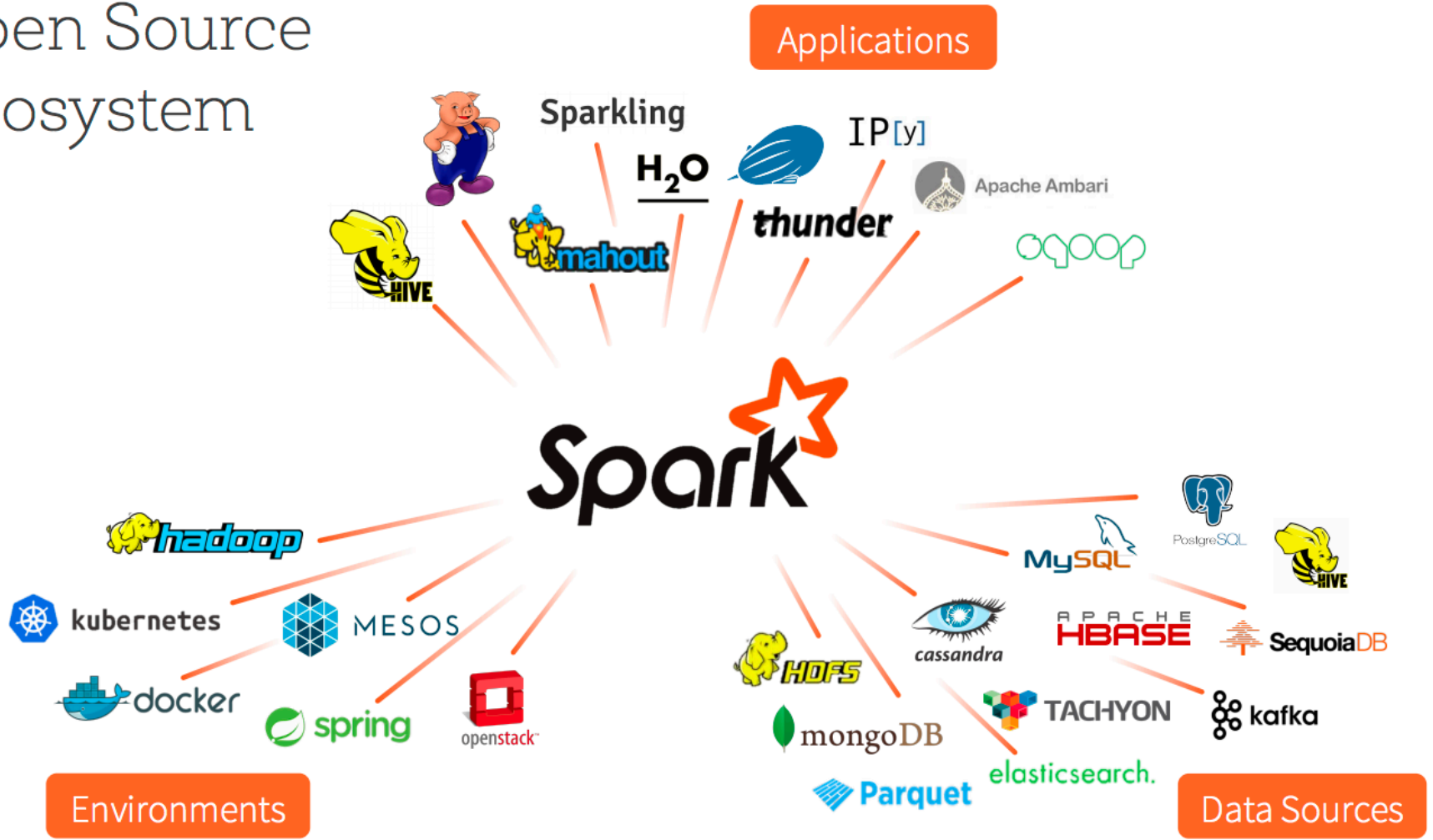
Spark 2.0


open source data processing engine built around
speed, ease of use, and sophisticated analytics

スピード、使いやすさ、洗練された分析を兼ね合わせた
オープンソースのデータ処理エンジン


Spark SQL + DataFrames
Streaming
MLlib
Machine Learning
GraphX
Graph Computation
Spark Core API
R
SQL
Python
Scala
Java

# Open Source Ecosystem



# About Databricks　Databricks について

Founded by creators of Spark & behind Spark development

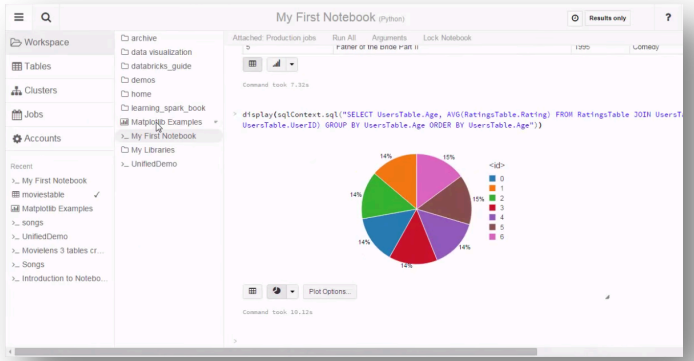
Spark開発者とSpark開発を支持する人たちによって設立された

Cloud Enterprise Spark Platform

- Cluster management, interactive notebooks, dashboards, production jobs, data governance, security, …

エンタープライズクラウド Spark プラットフォーム

・クラスタ管理、対話型ノートブック
・ダッシュボード、ジョブ生成
・データガバナンス、セキュリティ




databricks

# 2015: Great Year for Spark　2015: Sparkにとって大きな年

Most active open source project in data (1000+ contributors)

データ上、最も活発なオープンソースプロジェクト (1000人以上の貢献者)

New language: R

新しい言語：R

Widespread industry support & adoption

幅広い業界サポートと採用

# Meetup Groups: December 2014


source: meetup.com

# Meetup Groups: December 2015


source: meetup.com

News room > News releases >

# IBM Announces Major Commitment to Advance Apache®Spark™, Calling it Potentially the Most Significant Open S Project of the Next Decade

IBM Joins Spark Community, Plans

S

Select a topic or

News releases

Related XML feeds

IBMはApache Spark の高度化へのコミットメントをアナウンス、次の10年で最も重要なオープンソースプロジェクトとなる可能性を秘めているという

**ARMONK, NY - 15 Jun 2015:** IBM (NYSE Apache®Spark™, potentially the most im being defined by data. At the core of this leading Analytics and Commerce platform will also put more than 3,500 IBM researc at more than a dozen labs worldwide; dor technology to the Spark open source ecos scientists and data engineers on Spark.

databricks

Forbes / Tech

5 Billionaire-Owned Energy Stocks With Triple Digit Potential

JUN 22, 2015 @ 10:28 AM 30,975 VIEWS

## Spark Or Hadoop -- Which Is The Best Big Data Framework?

Spark Or Hadoop – どちらがベストなビッグデータフレームワーク？

One qu framew are not

Spark has overtaken Hadoop as the most active open source Big Data project. While they

TIME INC. NETWORK : FORTUNE MONEY TIME SPORTS ILLUSTRATED GOLF TRAVEL + LEISURE PEOPLE MORE

SUBSCRIBE

FORTUNE 500

TECH BIG DATA

## Survey shows huge popularity spike for Apache Spark

by Derrick

Apache Spark が人気急上昇だと調査結果が示す


article explaining the essential
tried to keep it accessible to

of the most popular tools used

work but recently the newer and
APA +2.63% Software Foundation

t mutually exclusive, as they are
mes faster than Hadoop in
system.

Oil prices fall on slowing global economic growth outlook SEPTEMBER 27, 2015

India's Modi grew emotional at Facebook as he recalls his childhood SEPTEMBER 27, 2015

September jobs, Congress' deadline, and Trevor Noah — 5 things to know this week SEPTEMBER 27, 2015

ISIS has a money problem SEPTEMBER 27, 2015

Indian Prime Minister tours Tesla factory, talks batteries & solar with Elon Musk SEPTEMBER 27, 2015

How CST Brands' Kim Lubel is dealing with her biggest fear SEPTEMBER 27, 2015

Here's how Bono and Mark Zuckerberg aim to bring the Internet to the globe SEPTEMBER 27, 2015

“Spark is the Taylor Swift of big data software.”

- Derrick Harris, Fortune

「Spark はビッグデータソフトウェアのテイラー・スウィフト※だ」
- ダーリック・ハリス、フォーチュン

※ 訳注：日本ではテレビ番組「テラスハウス」の主題歌などで有名

"Spark is the **北川景子**
of big data software."

(A Japanese engineer told me)

databricks™

# How are people using Spark?

# Diverse Runtime Environments

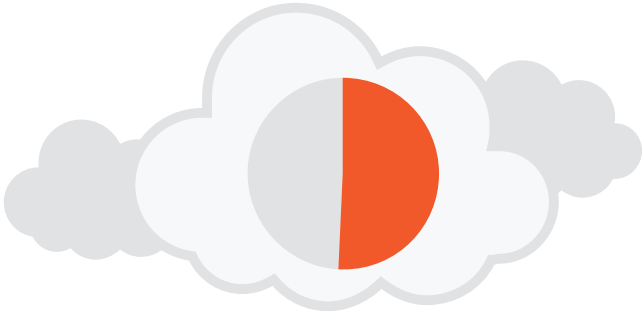
## さまざまな実行環境

Cluster Managers

**48%** Standalone mode

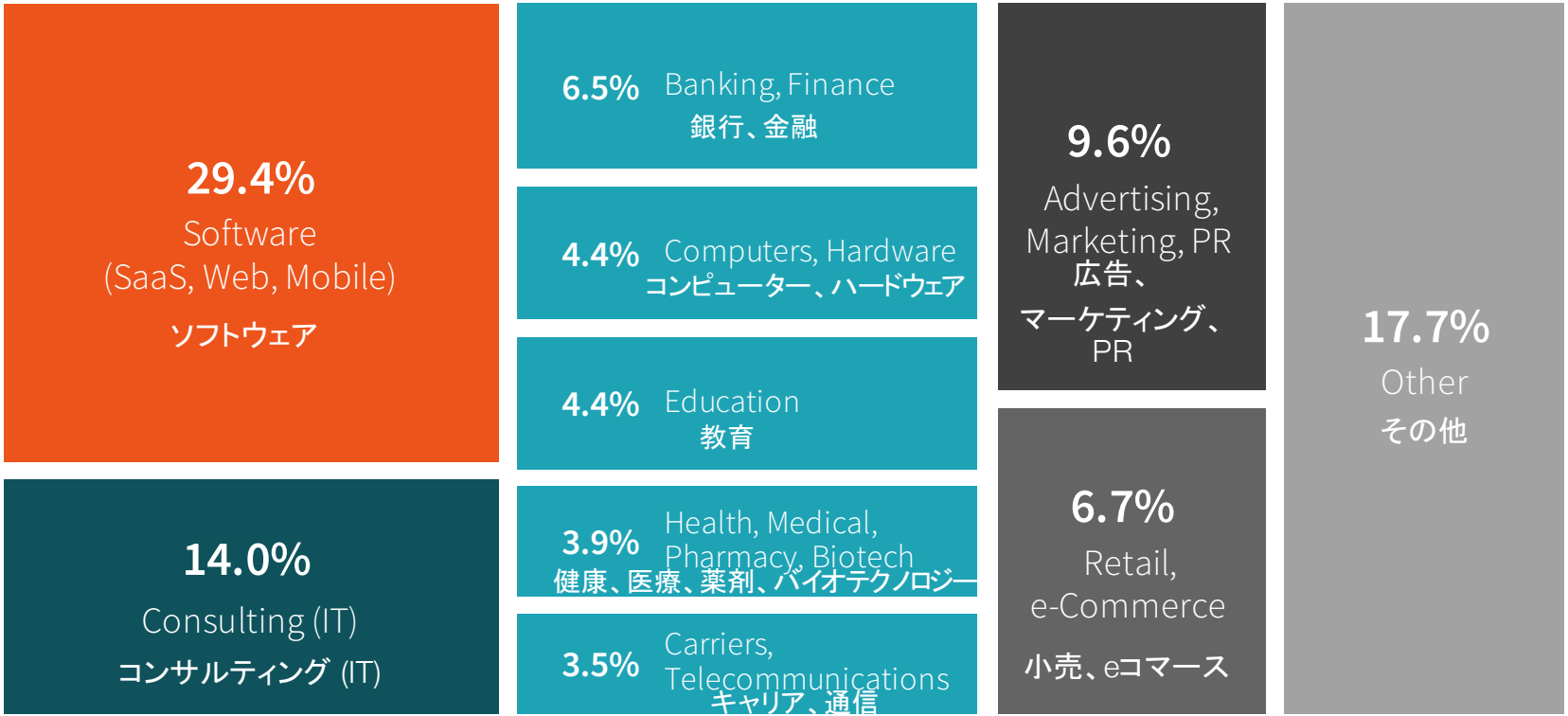
**40%** YARN

**11%** Mesos



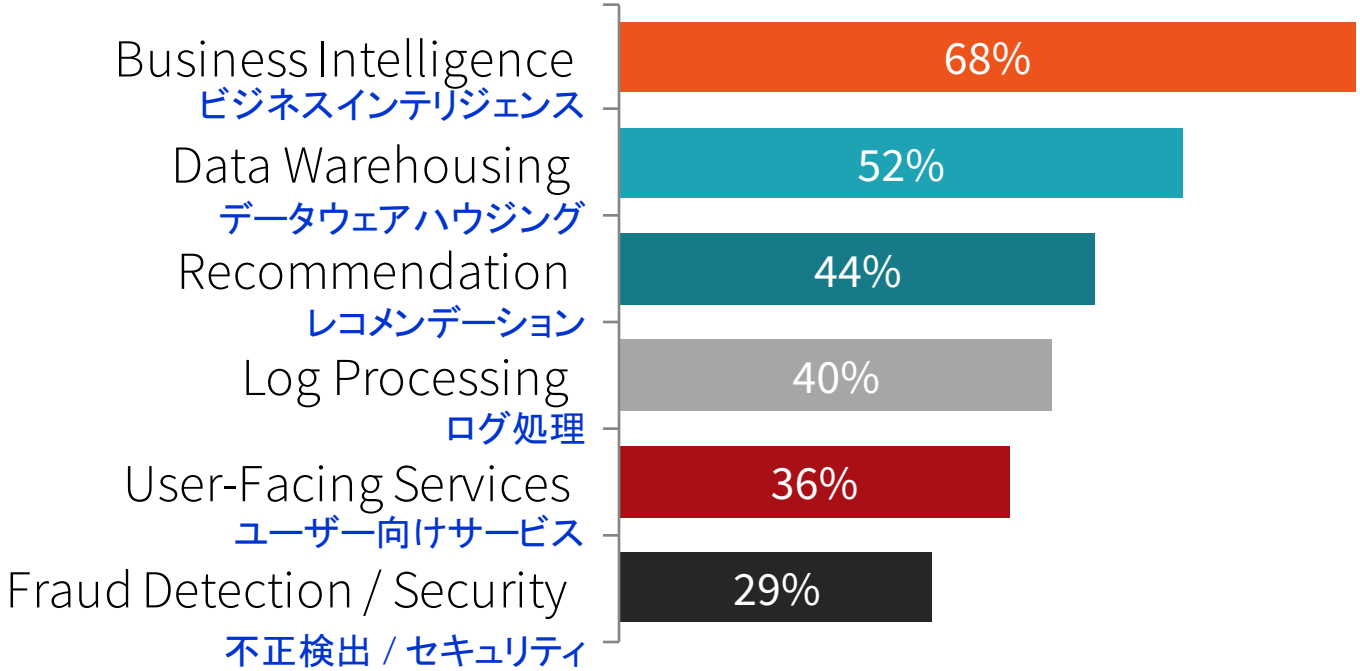
**51%** on a public cloud

# Industries Using Spark Sparkを利用している業界




databricks

# Top Applications 上位のアプリケーション

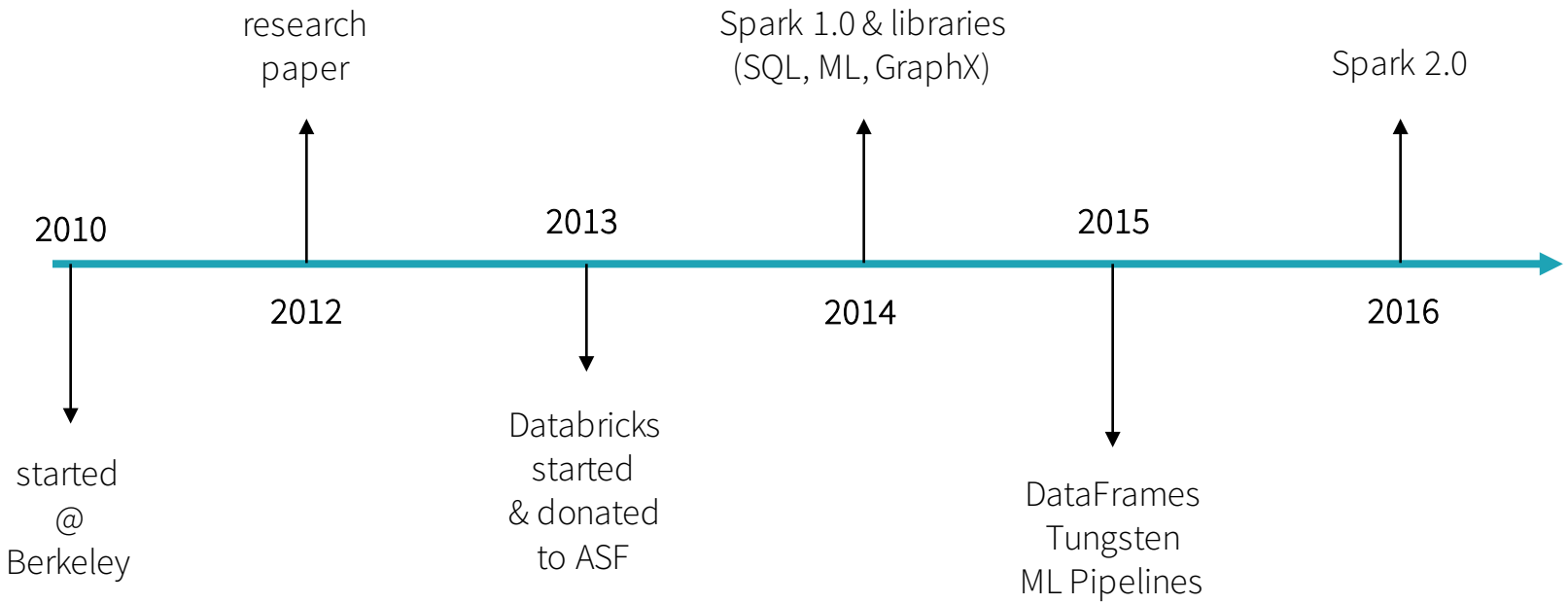


Are we done?

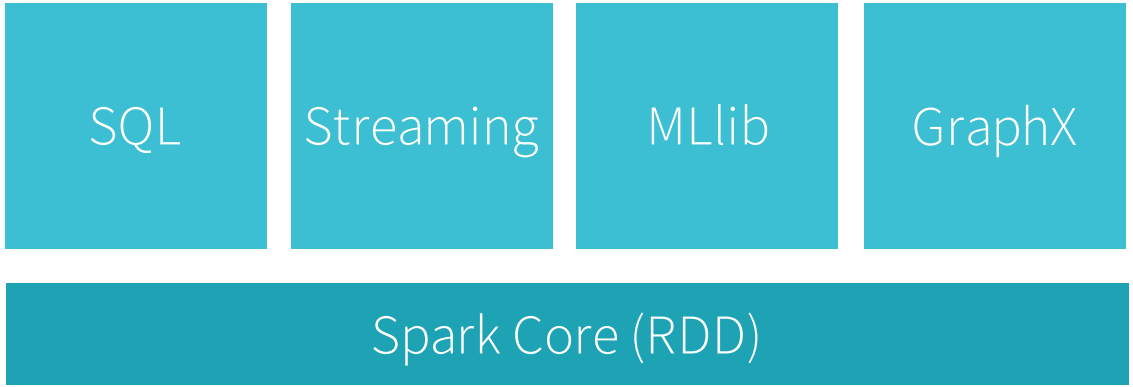
もう完成？

No. Development is faster than ever!

いいえ。開発は今まで以上に活発になって続いている！

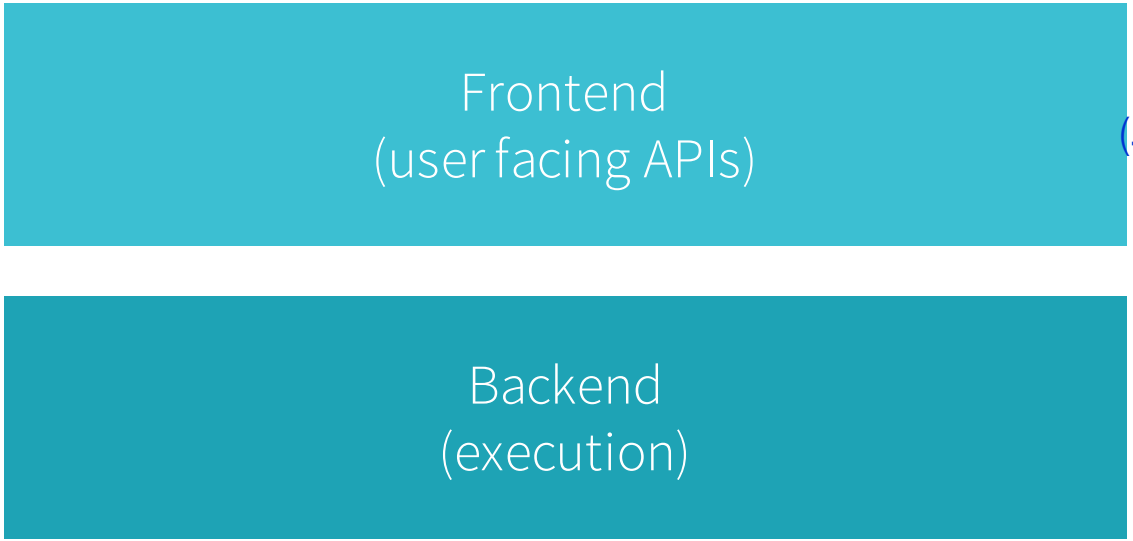

research
paper
Spark 1.0 & libraries
(SQL, ML, GraphX)
Spark 2.0
2010
2012
2013
2014
2015
2016
started
@
Berkeley
Databricks
started
& donated
to ASF
DataFrames
Tungsten
ML Pipelines

# Spark stack diagram Sparkのスタック図



# Spark stack diagram (a different take)

Sparkのスタック図 (違う見方で)

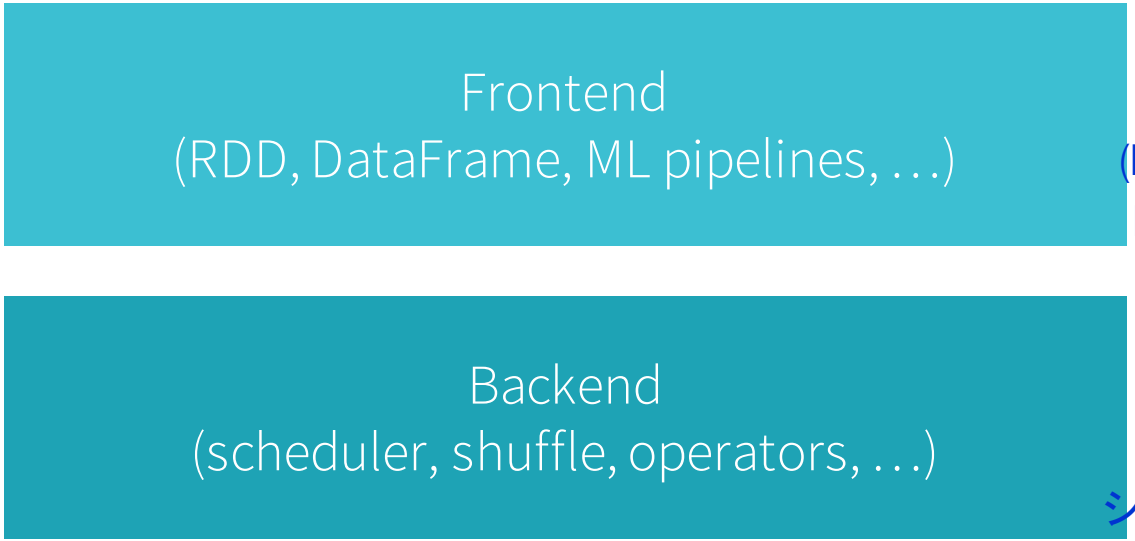



フロントエンド (ユーザーに面するAPI)

バックエンド (実行)

# Spark stack diagram (a different take)

# Sparkのスタック図 (違う見方で)




フロントエンド
(RDD, DataFrame, ML pipelines, …)

バックエンド
(スケジューラ、シャッフル、演算子、…)

# Spark 2.0

Frontend
API Foundation
フロントエンド API の創設

Streaming
DataFrame/Dataset
SQL
ストリーミング
DataFrame/Dataset
SQL

Backend
10X Performance
バックエンド 10倍のパフォーマンス

Whole-stage Codegen
Vectorization
全ステージ コード生成
ベクトル化

# Guiding Principles for API Foundation

API を創るにあたっての指針

1. Simple yet expressive

   シンプルだが表現豊かに

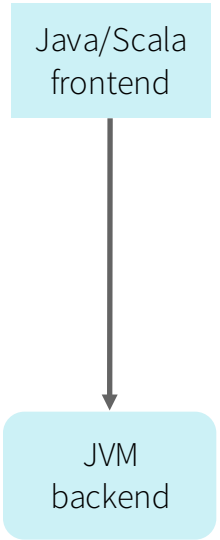
2. (Semantics) well-defined

   (セマンティクスが) 十分定義されている

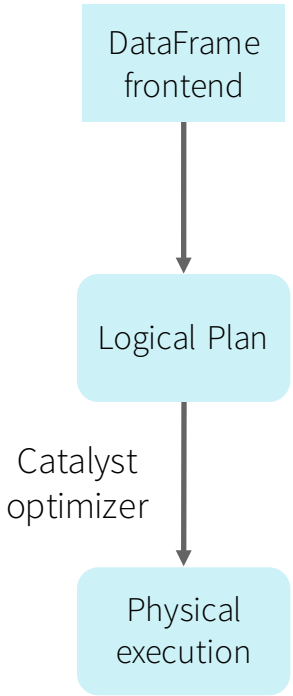
3. Sufficiently abstracted to allow optimized backends

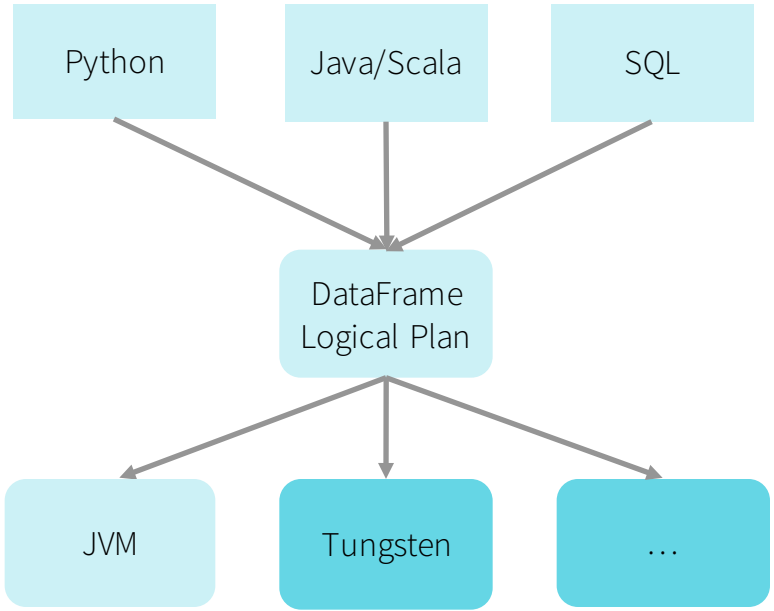
   バックエンドの最適化ができるよう十分に抽象化されている

RDD




DataFrame




Python
Java/Scala
SQL
DataFrame
Logical Plan
JVM
Tungsten
…

# API Foundations in Spark 2.0

Spark 2.0 におけるAPIの創設

1. Streaming DataFrames

   ストリーミング DataFrames

2. Maturing and merging DataFrame and Dataset

   DataFrame と Dataset の成熟とマージ

3. ANSI SQL

   - natural join, subquery, view support

     自然結合、サブクエリ、ビューのサポート

# Challenges with Stream Processing

ストリーム処理に関する課題

Stream processing is hard to reason about

- Output over time
- Late data
- Failures
- Distribution

ストリーム処理が難しい理由は

・長い期間に渡るアウトプット
・遅れてくるデータ
・障害
・分散

And all this has to work across complex operations

- Windows, sessions, aggregation, etc

これらすべてが複雑なオペレーションにわたって機能しなければならない

・ウィンドウ、セッション、アグリゲーション、など

# Next-gen Streaming with DataFrames

DataFramesによる次世代ストリーミング

1. Easy-to-use APIs (batch, streaming, and interactive)

1. 使いやすいAPI (バッチ、ストリーミング、インタラクティブ)

2. Well-defined semantics

- Out-of-order data
- Failures
- Sources/sinks with exactly-once semantics

2. うまく定義されたセマンティクス

・順序通りでないデータ

・障害

・exactly-once セマンティクスを持つ source / sink

3. Leverages Tungsten backend

3. Tungsten バックエンドの利用

# Next-gen Streaming with DataFrames

## DataFramesによる次世代ストリーミング

1. Easy-to-use APIs (batch, streaming, and interactive)
   1. 使いやすいAPI (バッチ、ストリーミング、インタラクティブ)
2. Well-defined semantics
   2. うまく定義されたセマンティクス
   - Out-of-order data
   - Failures
   - Sources/sinks with exactly-once semantics
   - 障害
   - exactly-once セマンティクスを持つ source / sink
3. Leverages Tungsten backend
   3. Tungsten バックエンドの利用

More details next few weeks

数週間後により詳細を

Spark is already pretty fast.

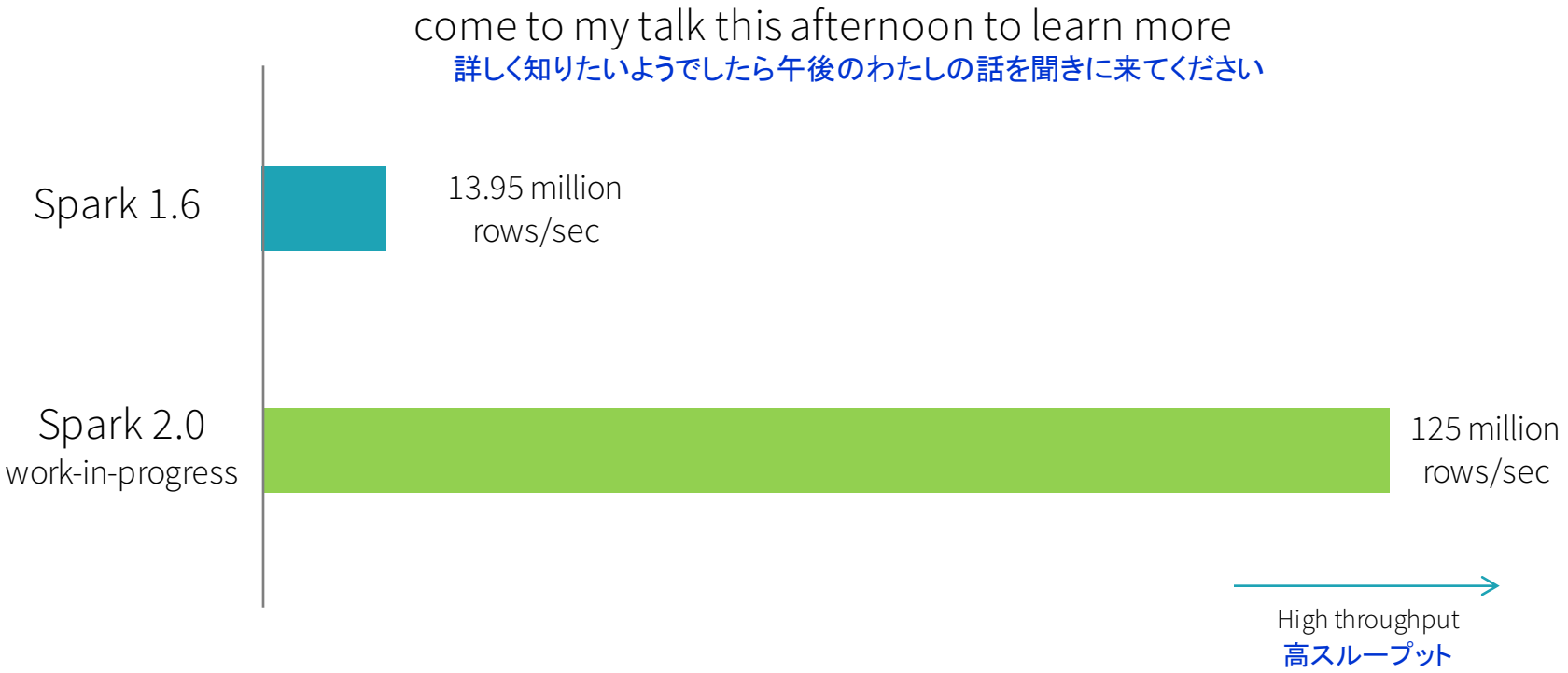
Spark はすでにかなり速い

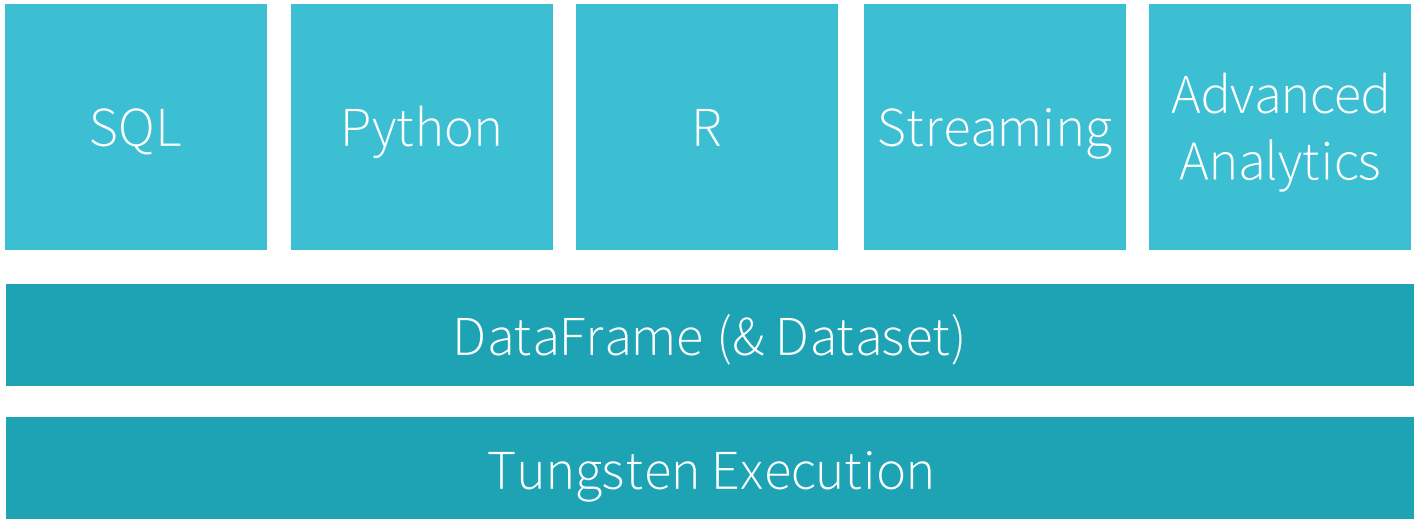
Can we make it 10X faster in 2.0?

2.0 で 10 倍高速にできるのだろうか？

# Teaser: SQL/DataFrame Performance

## 少しだけ宣伝: SQL/DataFrame のパフォーマンス




SQL
Python
R
Streaming
Advanced Analytics
DataFrame (& Dataset)
Tungsten Execution

# Spark 2.0 Release Schedule

Spark 2.0 のリリーススケジュール

Under active development on GitHub

GitHub 上で活発に開発中

March – April: code freeze

3月-4月：コードフリーズ

April – May: official release

4月-5月：正式リリース

# ありがとうございました

@rxin