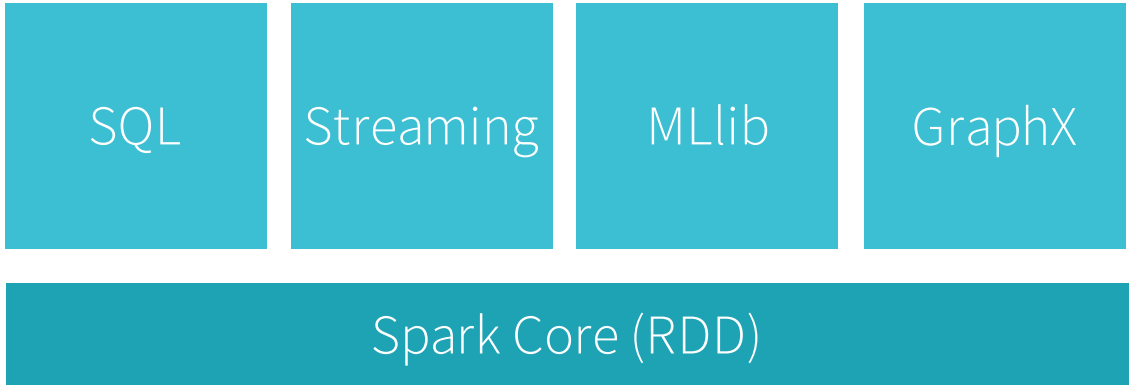
State of Spark, and where it is going

Reynold Xin @rxin

Strata Singapore

Dec 3rd, 2015

Spark stack diagram



A Great Year for Spark

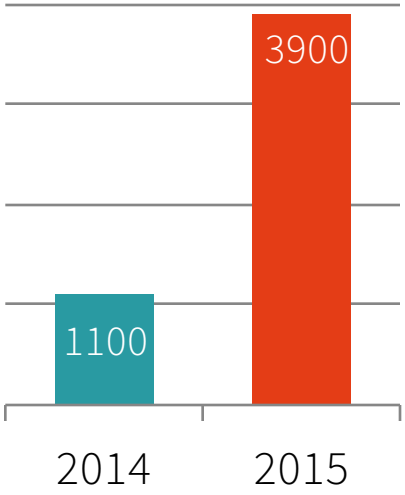
Most active open source project in big data

New language: R

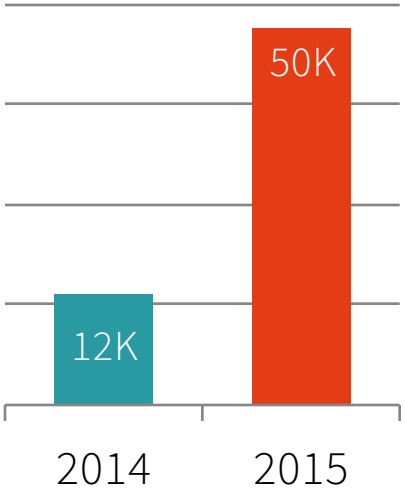
Widespread industry support & adoption

Community Growth

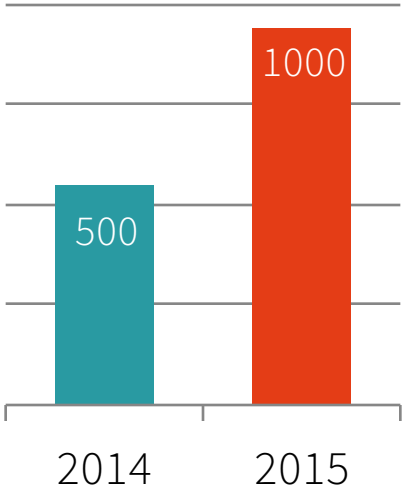
Summit Attendees



Meetup Members



Developers Contributing



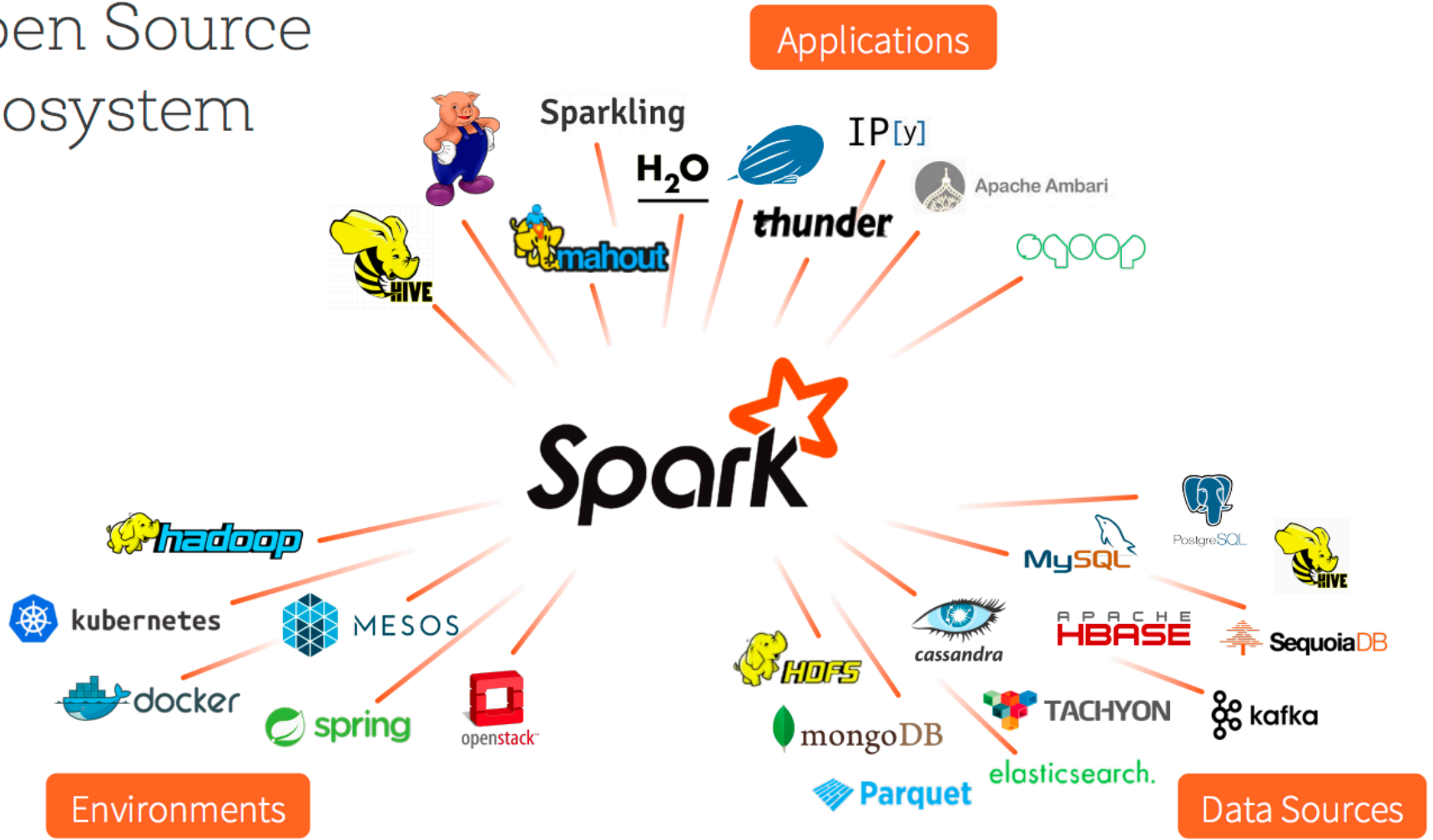
Meetup Groups: December 2014

source: meetup.com

Meetup Groups: December 2015

source: meetup.com

Open Source Ecosystem



Users

1000+ companies

. . .

Distributors + Apps

50+ companies

. . .

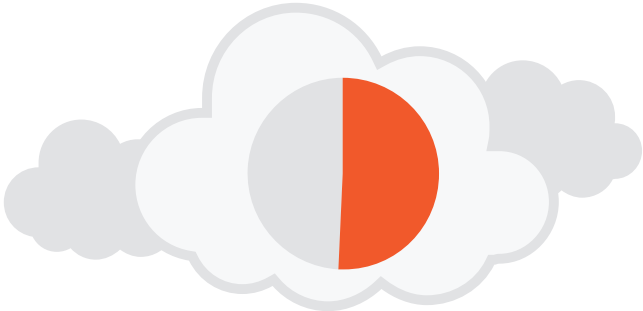
Diverse Runtime Environments

Cluster Managers

48% Standalone mode

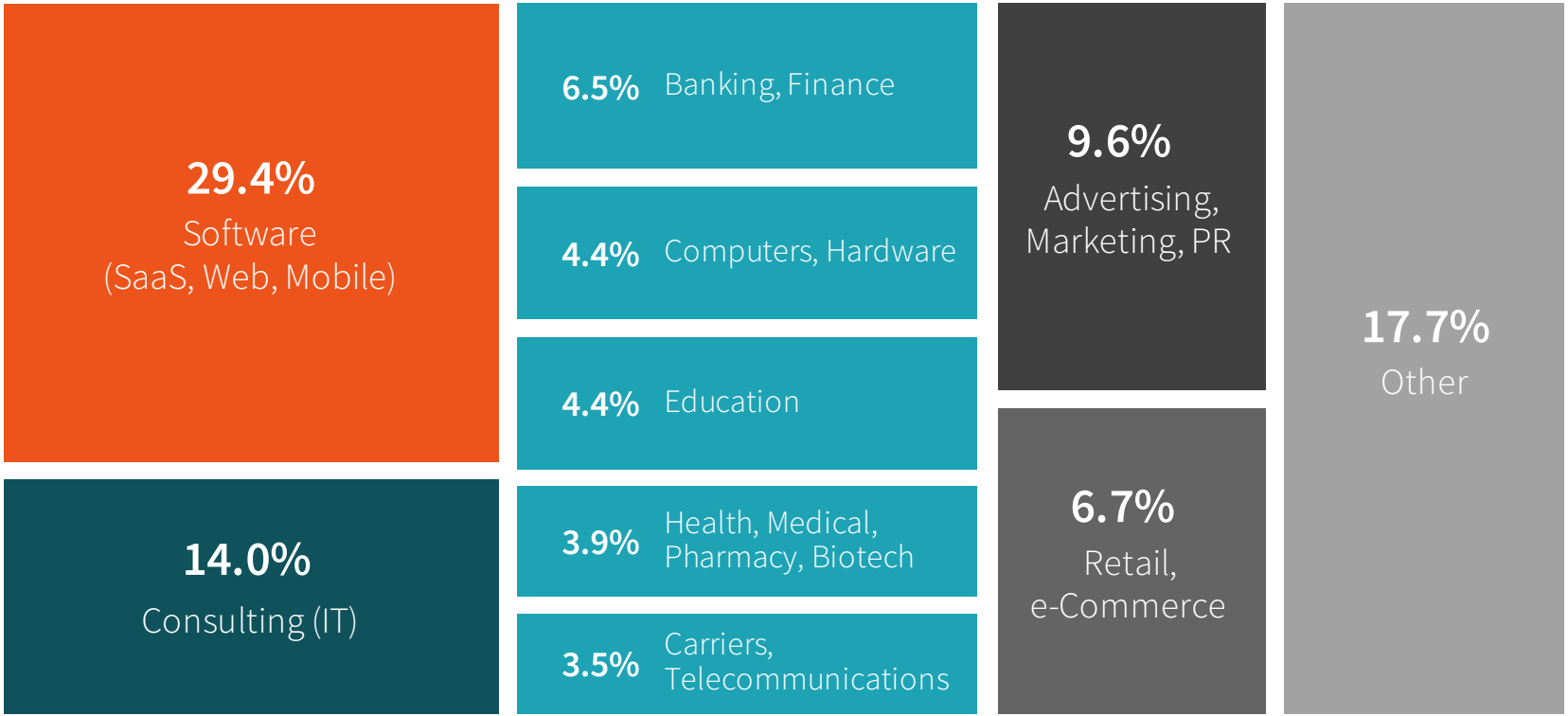
40% YARN

11% Mesos

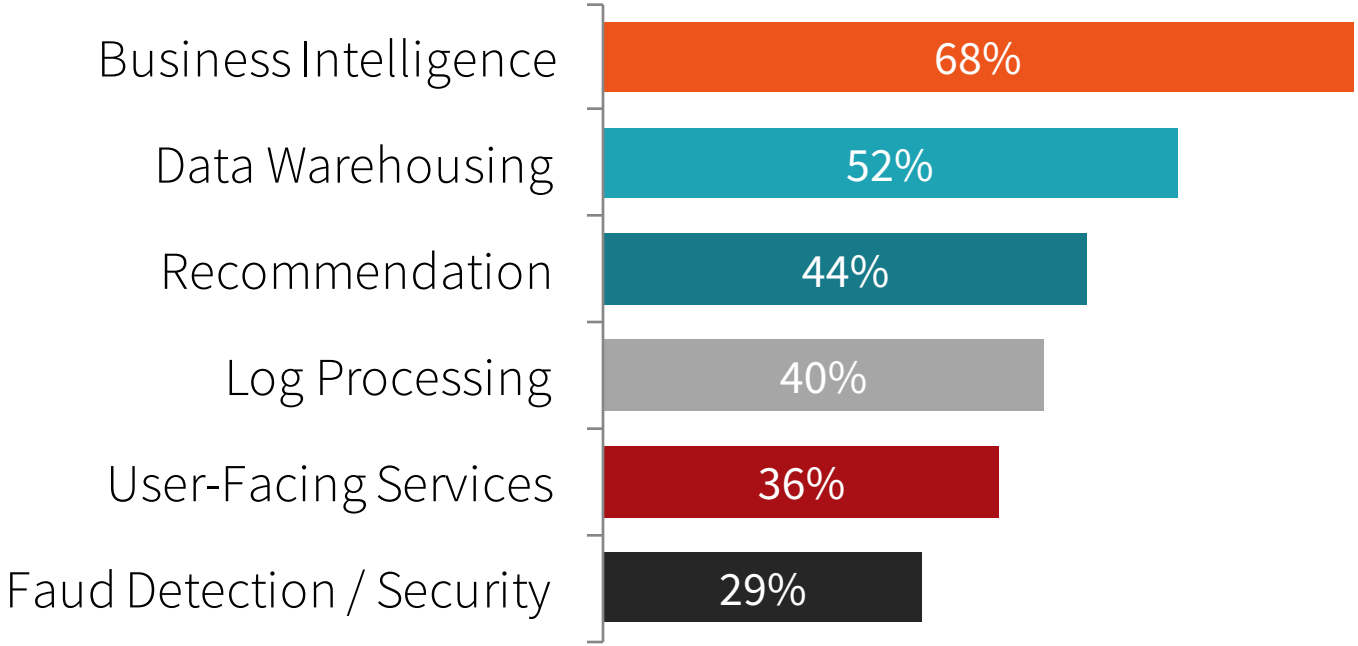


51% on a public cloud

Industries Using Spark



Top Applications

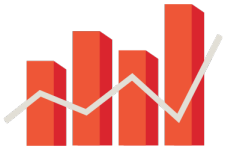


Largest Cluster & Daily Intake

800 million+
active users

8000+
nodes

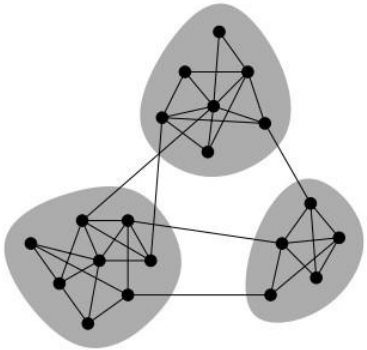
150 PB+
1 PB+/day



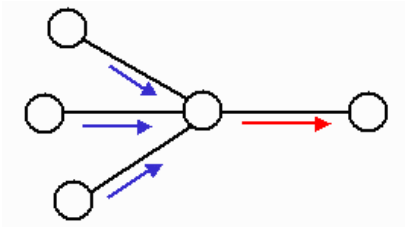
Tencent 腾讯 数据平台部

Alibaba Taobao

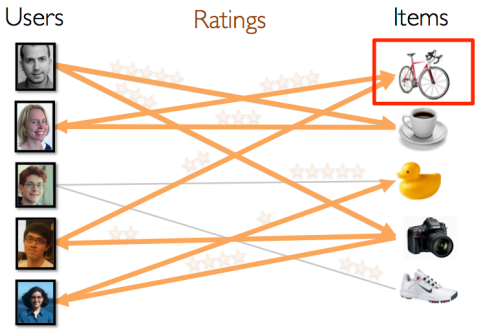
clustering
(community detection)



belief propagation
(influence & credibility)



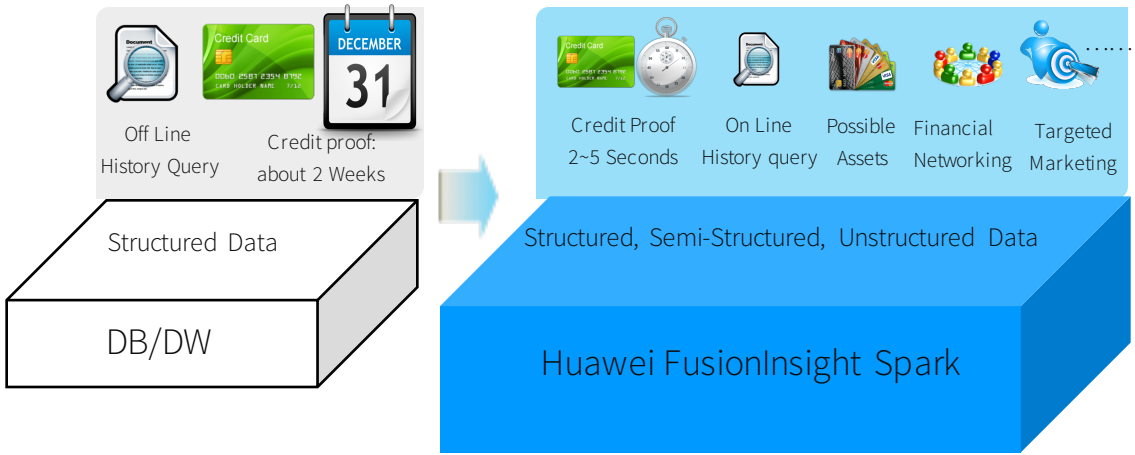
collaborative filtering
(recommendation)



Top Retail Bank & Huawei

Top Retail Bank

Huawei



History Query	1 year	↑7 years+
Micro- loan Conversion Rate	Higher	↑40X
Credit Proof	15days	↓2-5s

databricks

Are We Done?

No! Development is faster than ever. Expect Spark 2.0 in 2016.

Biggest technical change in 2015 was DataFrames

- Moves many computations onto the relational Spark SQL optimizer

Enables both new APIs and more optimization, which is now happening through Project Tungsten

Coming in Spark 1.6

Dataset API: typed interface over DataFrames / Tungsten

- Common ask from developers who saw DataFrames

```
case class Person(name: String, age: Int)

val dataframe = read.json("people.json")
    val ds: Dataset[Person] = dataframe.as[Person]

ds.filter(p => p.name.startsWith("M"))
      .groupBy("name")
      .avg("age")
```

Other Upcoming Features

DataFrame integration with GraphX and Streaming

More Tungsten features: faster in-memory cache, SSD storage, better code generation

Data sources for Streaming

Thank you.

@rxin