State of Spark, and where it is going

Reynold Xin 辛湜
@hashjoin

2015-12-10, Beijing BDTC

databricks™

About Databricks

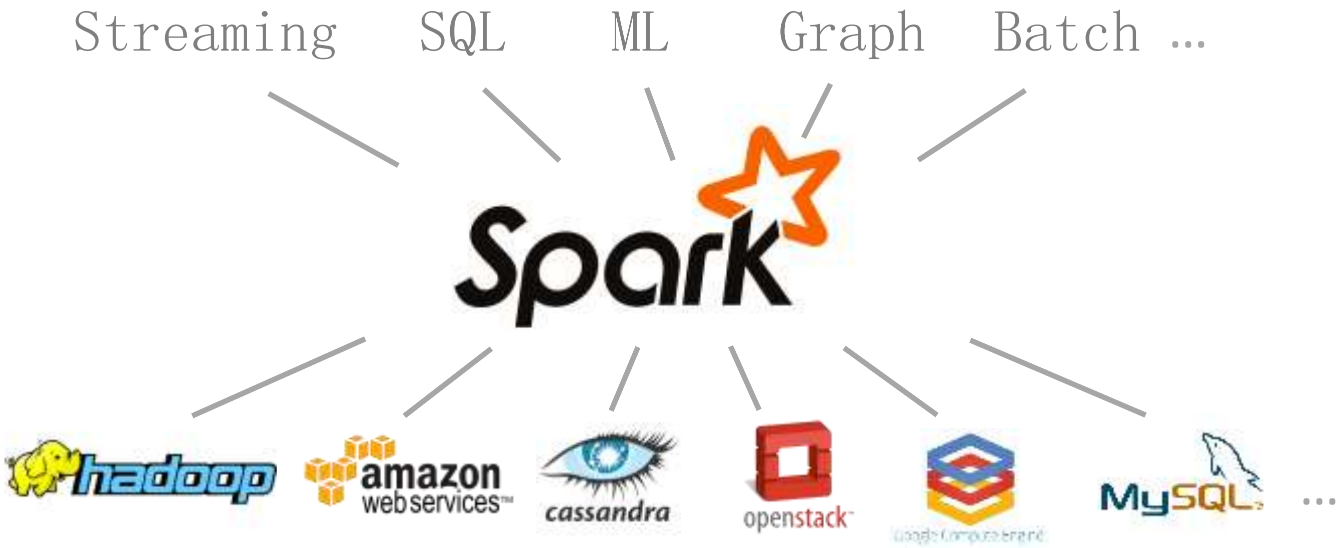
Founded by creators of Spark in 2013

Cloud service for end-to-end data processing

- Interactive notebooks, dashboards, production jobs, security, …

Our Goal for Spark

Unified engine across data workloads and platforms



Streaming SQL ML Graph Batch …

Spark "Hall of Fame"

★

LARGEST CLUSTER

Tencent
(8000+ nodes)

★

LARGEST SINGLE-DAY INTAKE

Tencent
(1PB+ /day)

★

LONGEST-RUNNING JOB

Alibaba
(1 week on 1PB+ data)

★

LARGEST SHUFFLE

Databricks PB Sort
(1PB)

★

MOST INTERESTING APP

Jeremy Freeman
Mapping the Brain at Scale
(with lasers!)

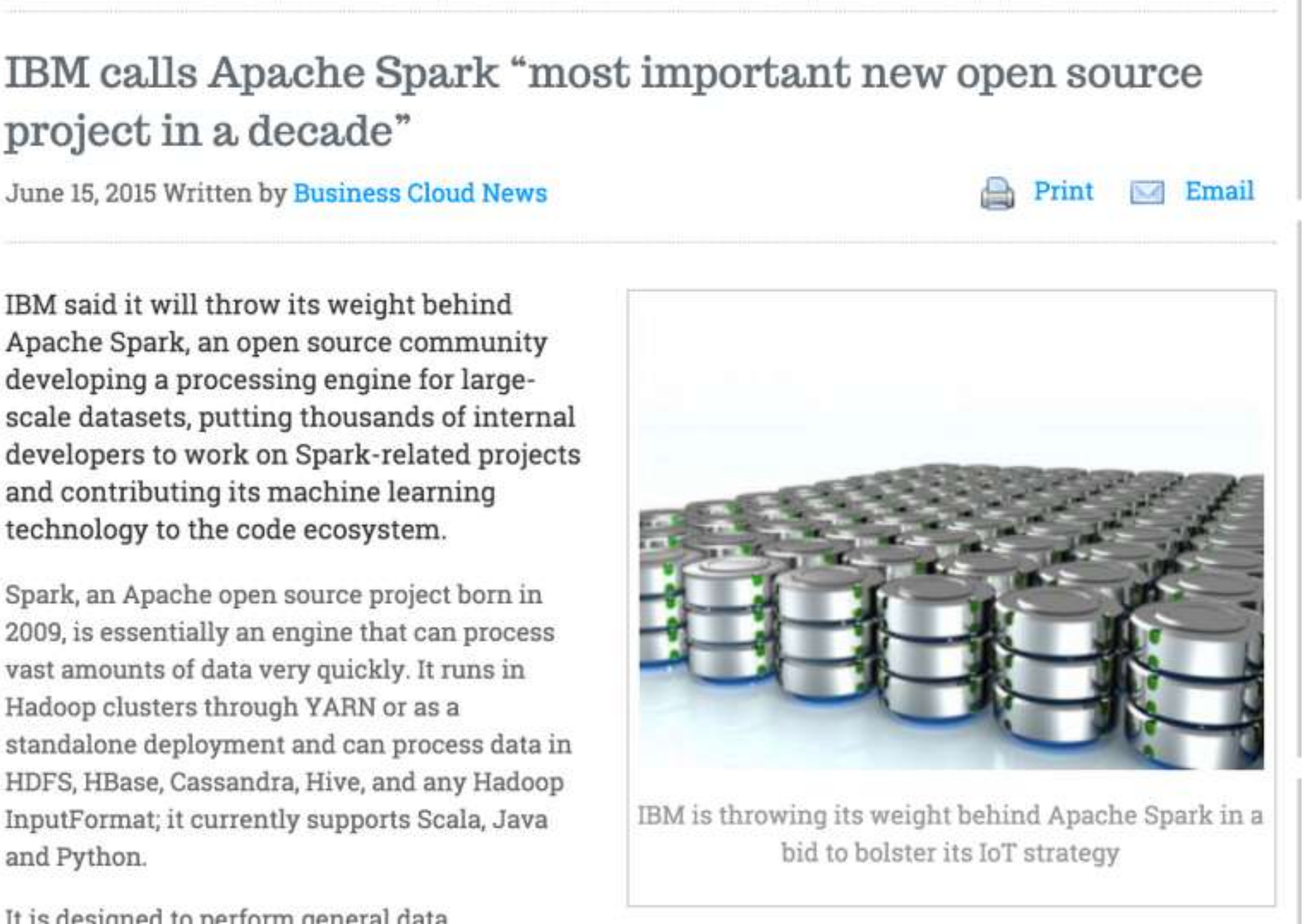
Based on Reynold's personal knowledge

A Great Year for Spark

Most active open source project in big data

New language: R

Widespread industry support & adoption



IBM calls Apache Spark "most important new open source project in a decade"

June 15, 2015 Written by Business Cloud News

IBM said it will throw its weight behind Apache Spark, an open source community developing a processing engine for large-scale datasets, putting thousands of internal developers to work on Spark-related projects and contributing its machine learning technology to the code ecosystem.

Spark, an Apache open source project born in 2009, is essentially an engine that can process vast amounts of data very quickly. It runs in Hadoop clusters through YARN or as a standalone deployment and can process data in HDFS, HBase, Cassandra, Hive, and any Hadoop InputFormat; it currently supports Scala, Java and Python.

IBM is throwing its weight behind Apache Spark in a bid to bolster its IoT strategy

It is designed to perform general data

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

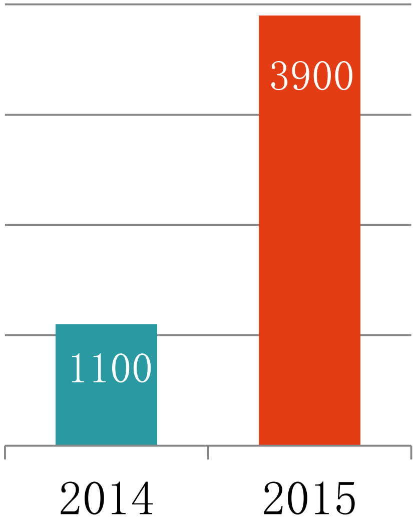
- Derrick Harris, Fortune

“Spark是大数据中的 Angelababy.”

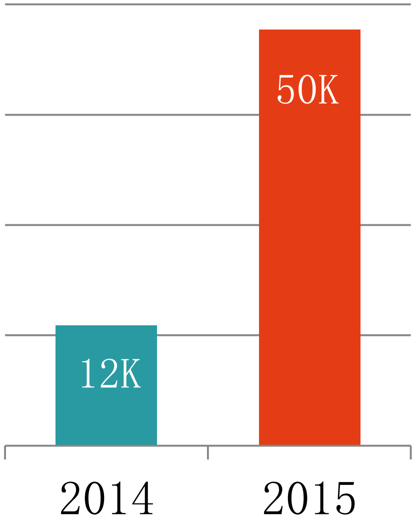
- Derrick Harris, Fortune

Community Growth

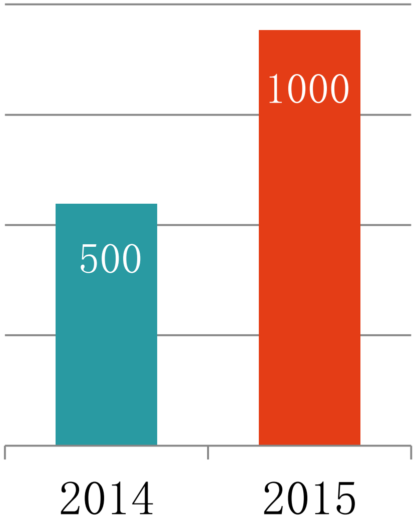
Summit Attendees



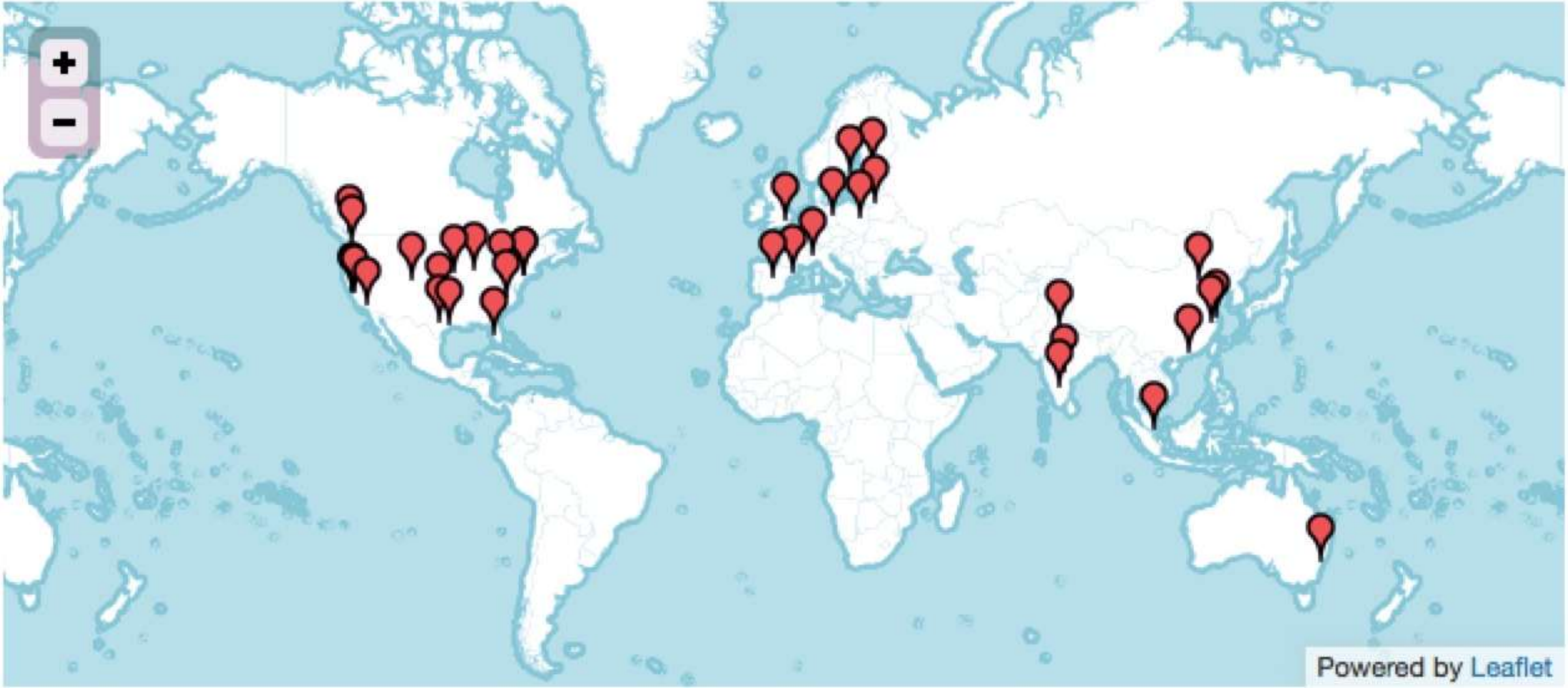
Meetup Members



Developers Contributing

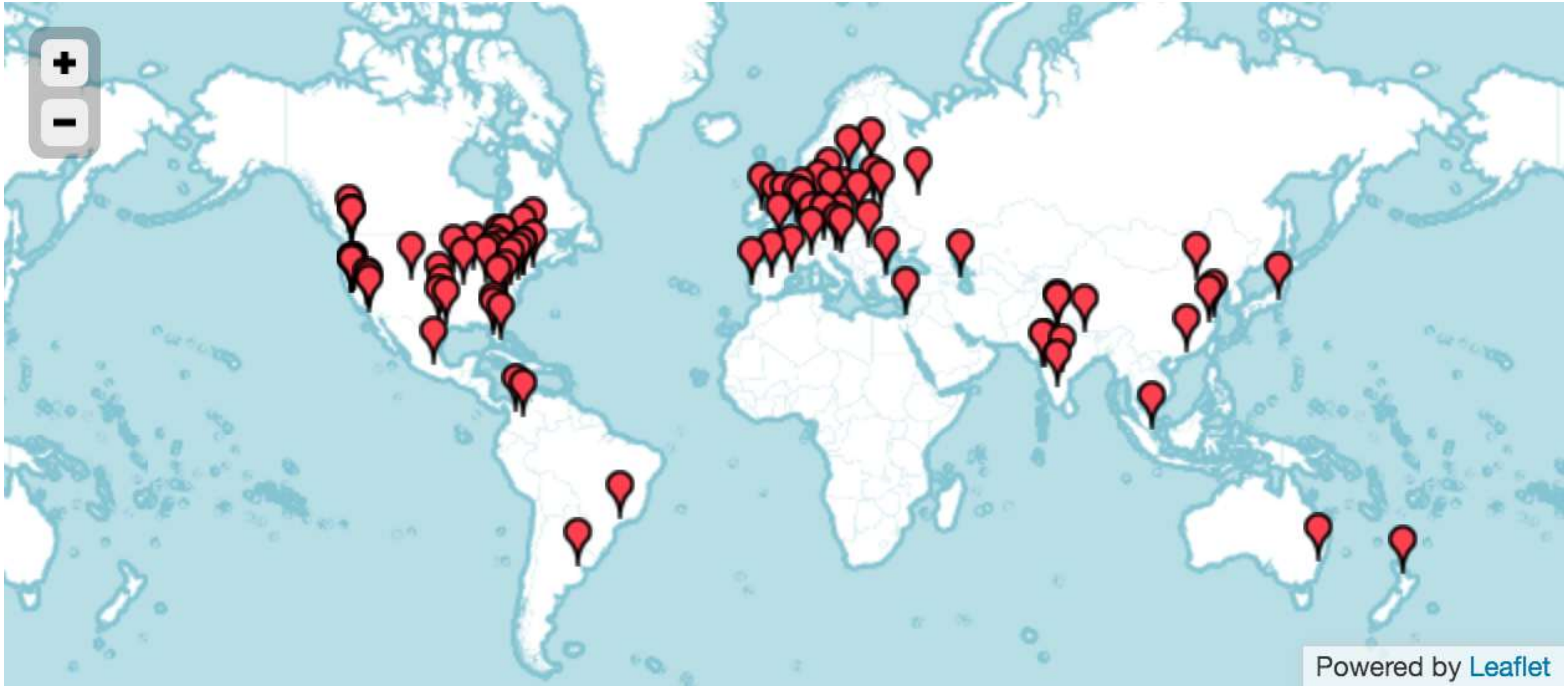


Meetup Groups: December 2014



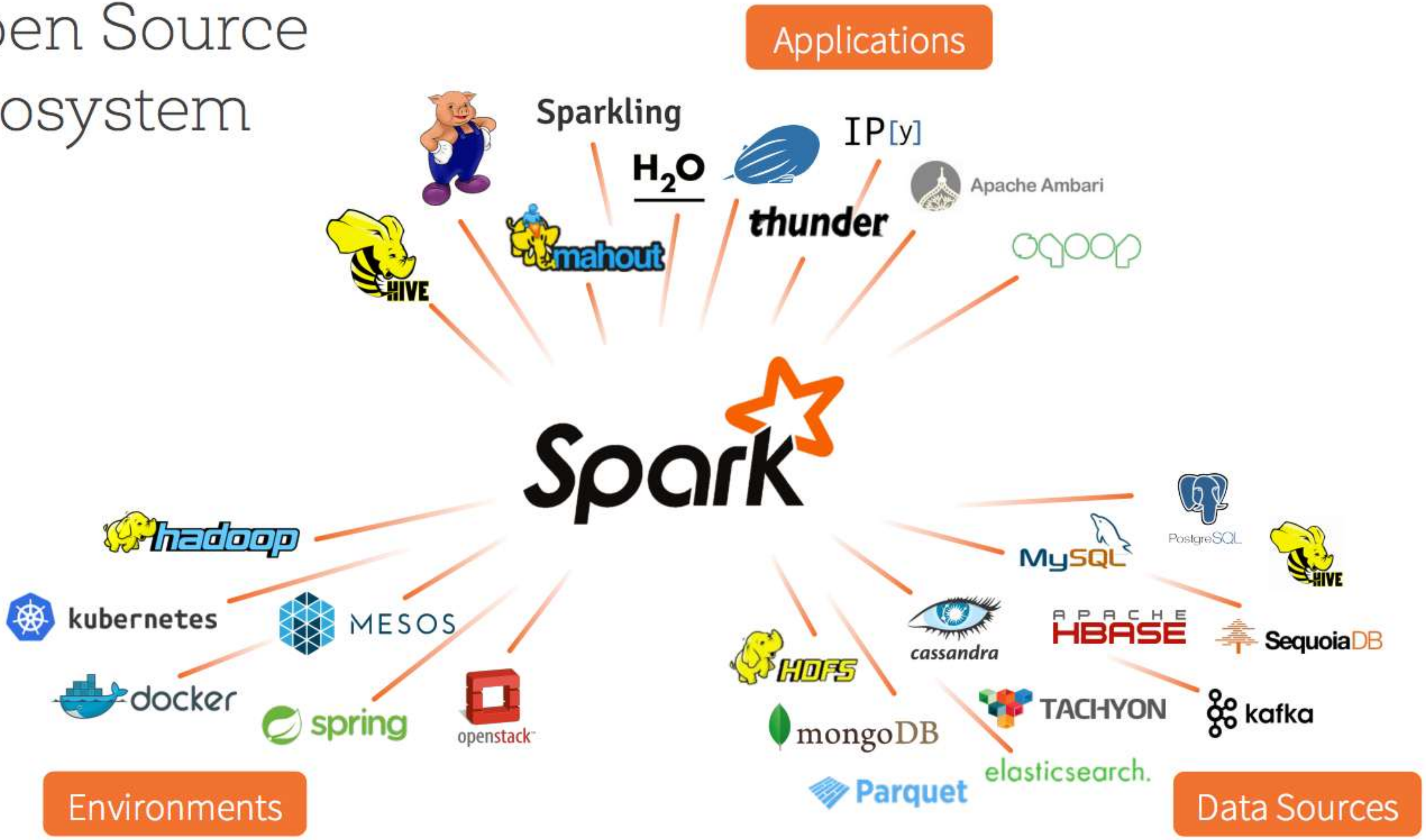
source: meetup.com

Meetup Groups: December 2015



source: meetup.com

Open Source Ecosystem



Users

1000+ companies

Goldman Sachs TOYOTA Tencent 腾讯 NOVARTIS

Baidu 百度 淘宝网 taobao.com JD.COM 京东

ebay UBER salesforce

airbnb Telefonica 大众点评 dianping.com

Palantir Microsoft NETFLIX

…

Distributors + Apps

50+ companies

databricks™

Hortonworks MAPR cloudera

SAP® IBM ORACLE

DATASTAX amazon web services™

+ a b l e a u Qlik alteryx

informatica MicroStrategy TIBCO®

…

Spark Survey 2015

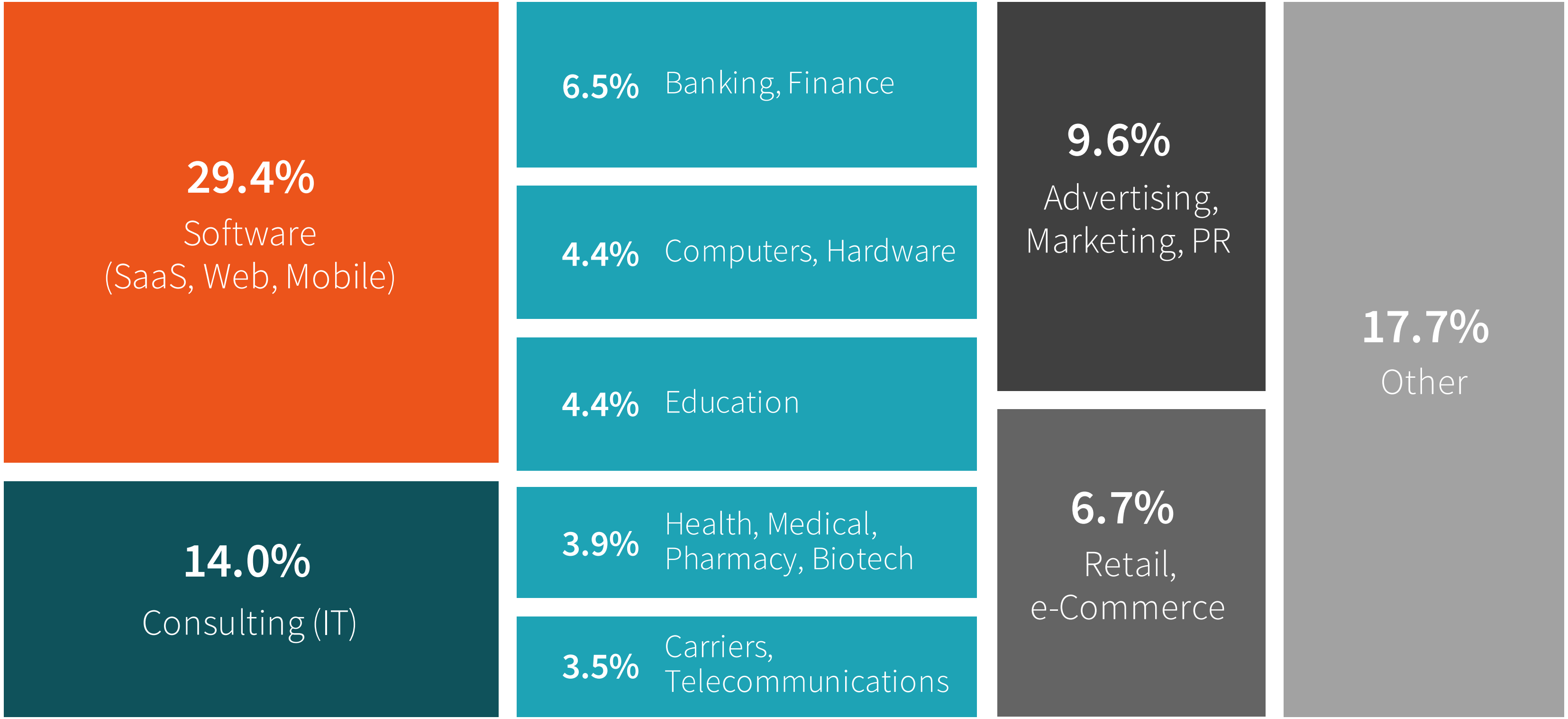
Databricks Survey

1400 respondents from 840 companies

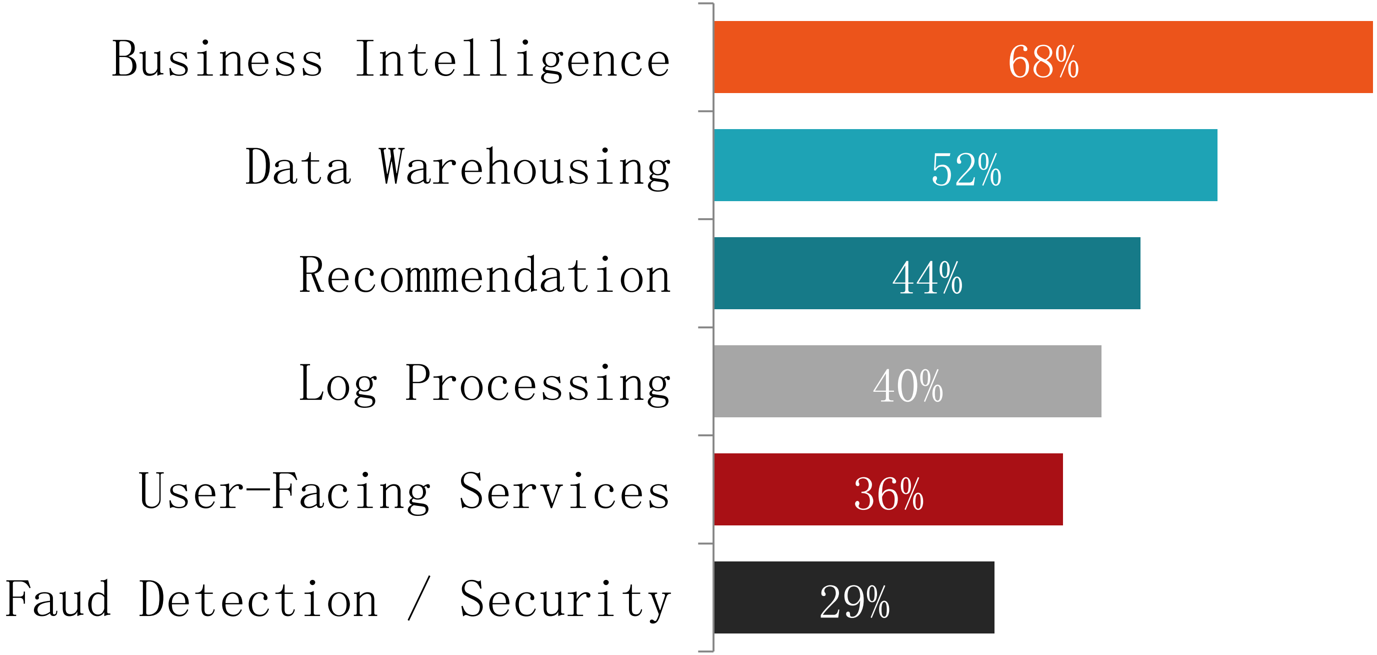
Three trends:

1) Diverse applications
2) More runtime environments
3) More types of users

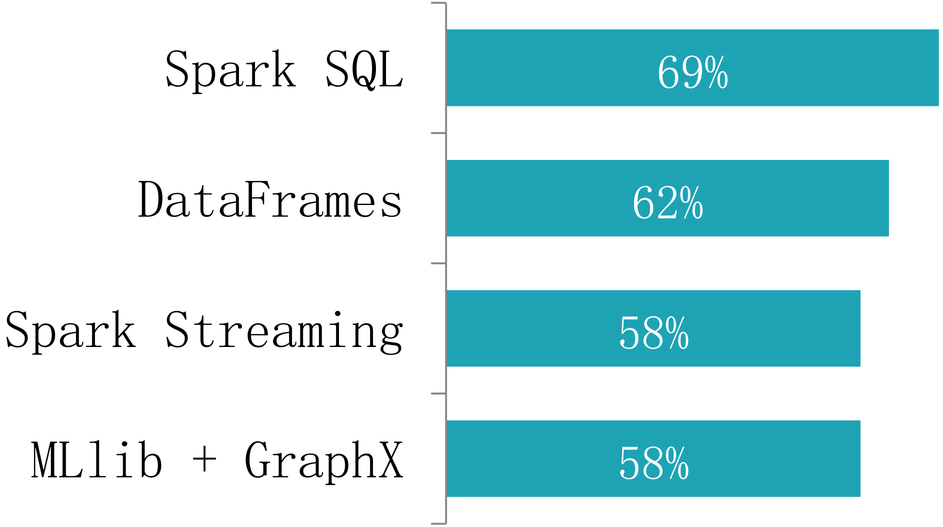
Industries Using Spark



Top Applications



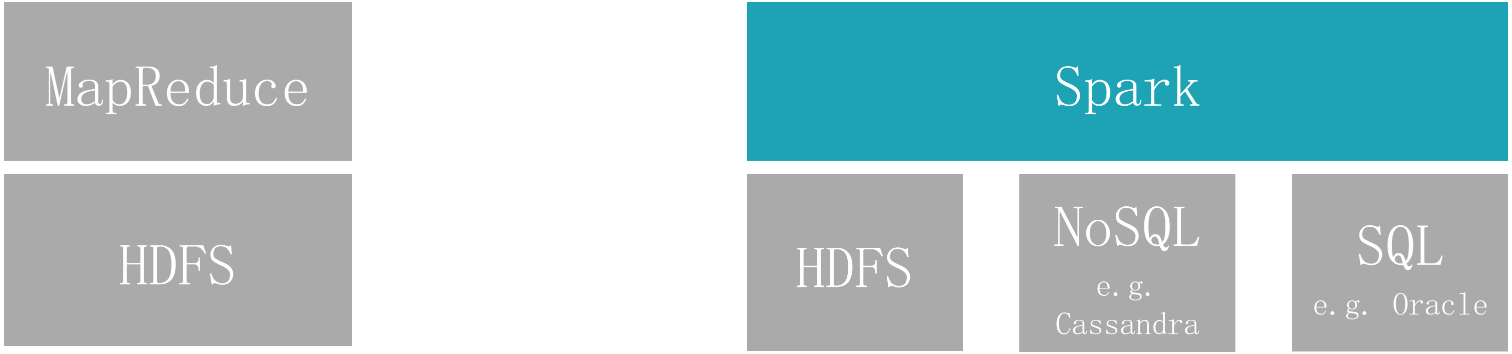
Spark Components Used



75%

of users use more than one component

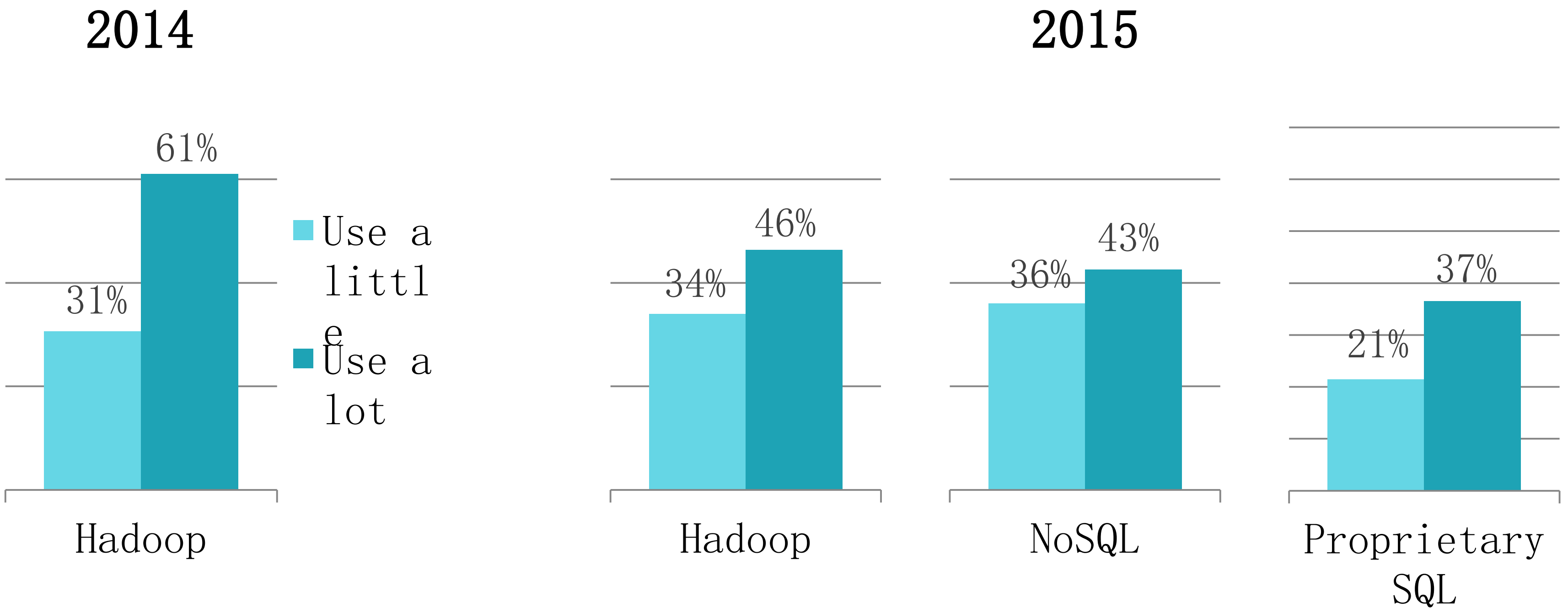
Diverse Runtime Environments



Hadoop: combined compute + storage

Spark: independent of storage layer

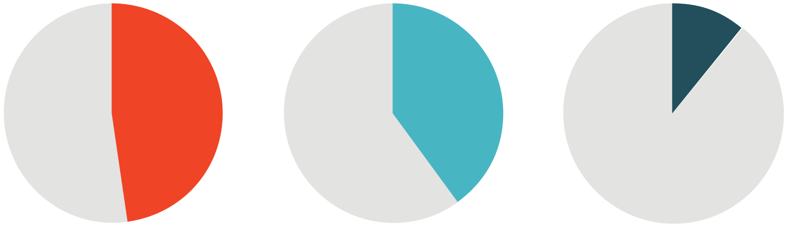
Diverse Runtime Environments



databricks

Diverse Runtime Environments

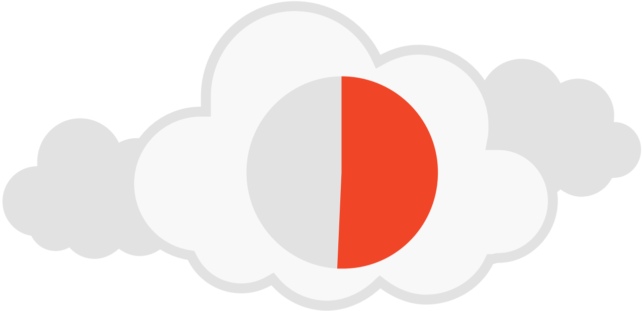
Cluster Managers



48% Standalone mode

40% YARN

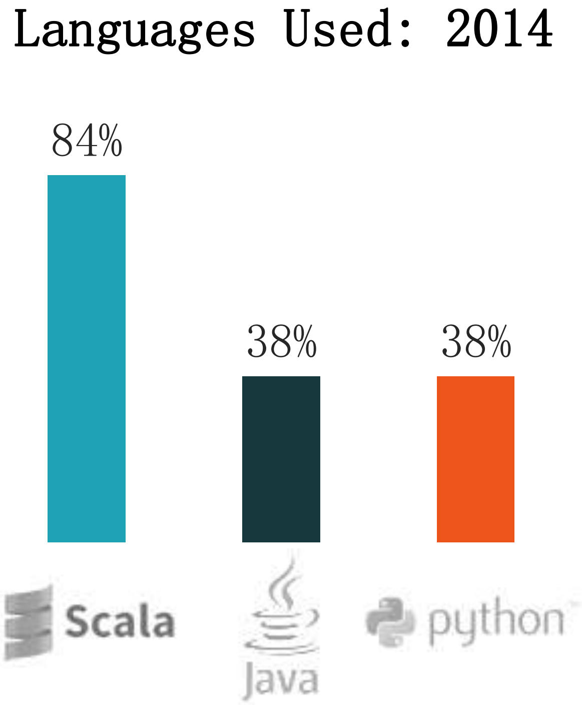
11% Mesos

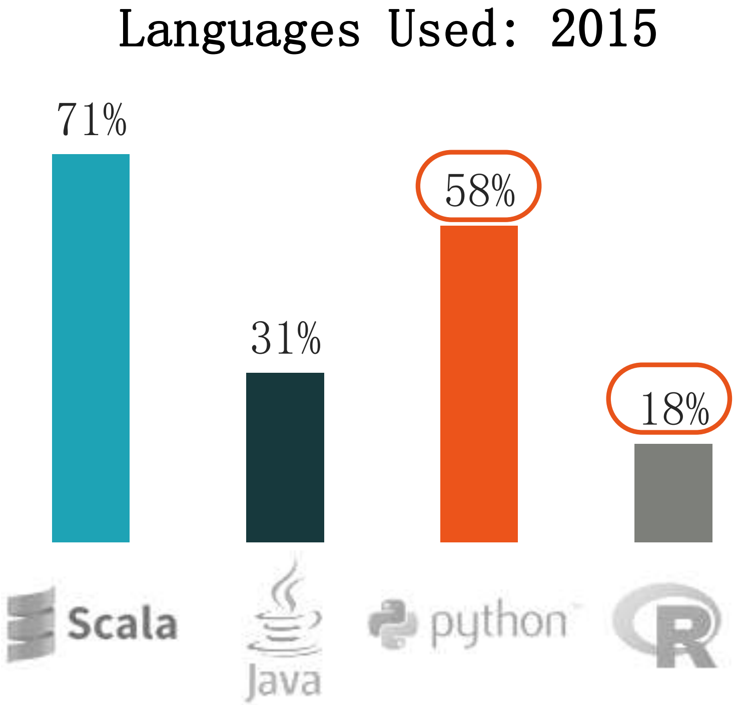


51% on a public cloud

databricks

Diversity of Users





Fastest Growing Components

+280%

increase in
Windows users

+56%

production use
of Streaming

+380%

production
use of SQL

Are We Done?

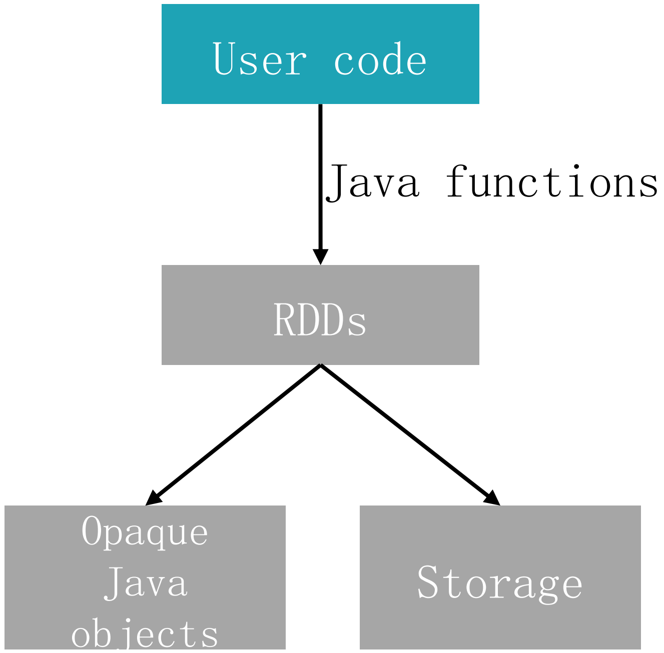
No! Development is faster than ever.

Biggest technical change in 2015 was DataFrames

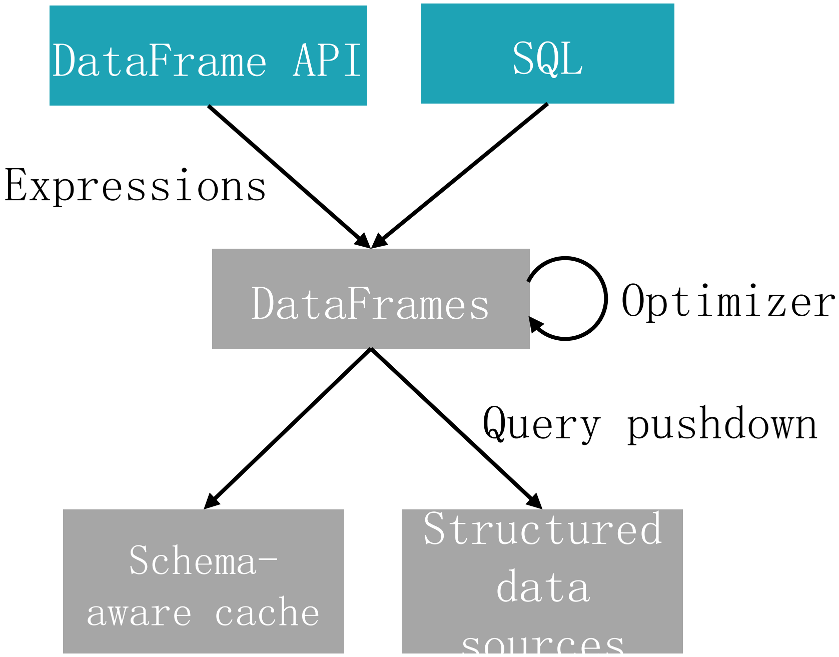
- Moves many computations onto the relational Spark SQL optimizer

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

Traditional Spark



DataFrames



databricks

3 Things to Look Forward To

Dataset API in Spark 1.6 (SPARK-9999)

Typed interface over DataFrames / Tungsten

```
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”)
```

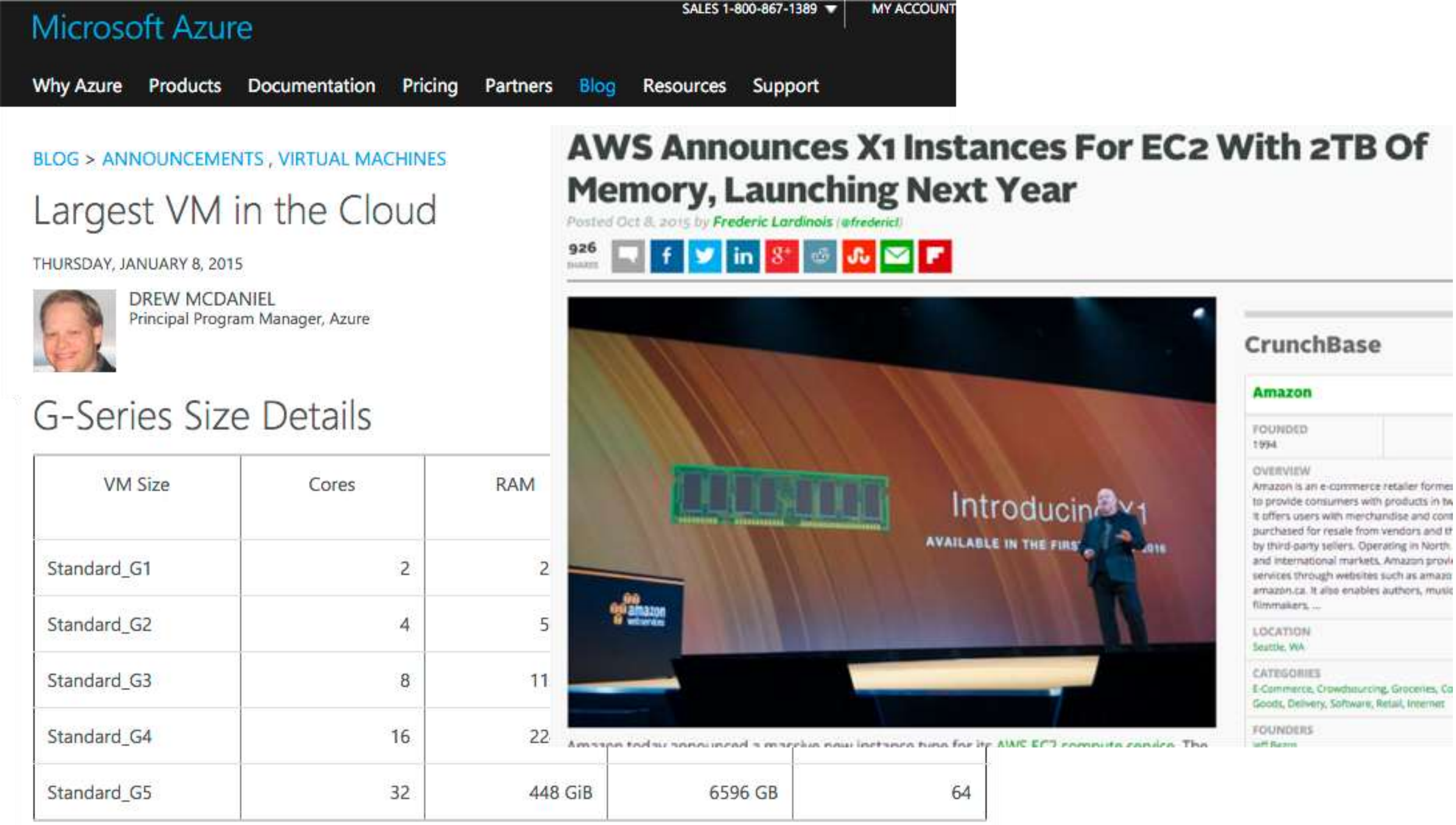
Streaming DataFrames

Easier-to-use APIs (batch, streaming, and interactive)

```
val stream = read.kafka("...")
stream.window(5 mins, 10 secs)
  .agg(sum("sales"))
  .write.jdbc("mysql://...")
```

And optimizations:

- Tungsten backends
- native support for out-of-order data
- data sources and sinks



BLOG > ANNOUNCEMENTS , VIRTUAL MACHINES

Largest VM in the Cloud

THURSDAY, JANUARY 8, 2015

DREW MCDANIEL
Principal Program Manager, Azure

G-Series Size Details

VM Size	Cores	RAM		
Standard_G1	2	2		
Standard_G2	4	5		
Standard_G3	8	11		
Standard_G4	16	22		
Standard_G5	32	448 GiB	6596 GB	64

AWS Announces X1 Instances For EC2 With 2TB Of Memory, Launching Next Year

Posted Oct 8, 2015 by Frederic Lardinois (@fredericl)

926 SHARES

Amazon today announced a massive new instance type for its AWS EC2 compute service. The

CrunchBase

Amazon

FOUNDED
1994

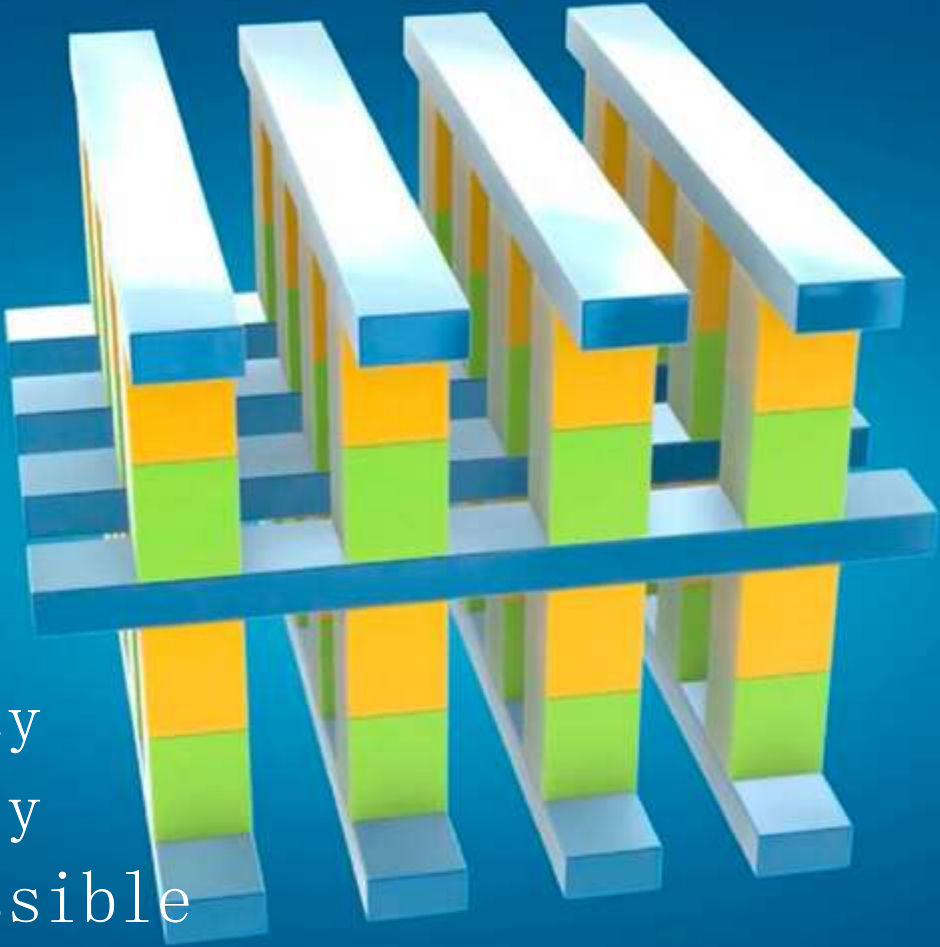
OVERVIEW
Amazon is an e-commerce retailer formed to provide consumers with products in tw
It offers users with merchandise and con
purchased for resale from vendors and th
by third-party sellers. Operating in North
and international markets, Amazon provi
services through websites such as amazo
amazon.ca. It also enables authors, music
filmmakers, ...

LOCATION
Seattle, WA

CATEGORIES
E-Commerce, Crowdsourcing, Groceries, Co
Goods, Delivery, Software, Retail, Internet

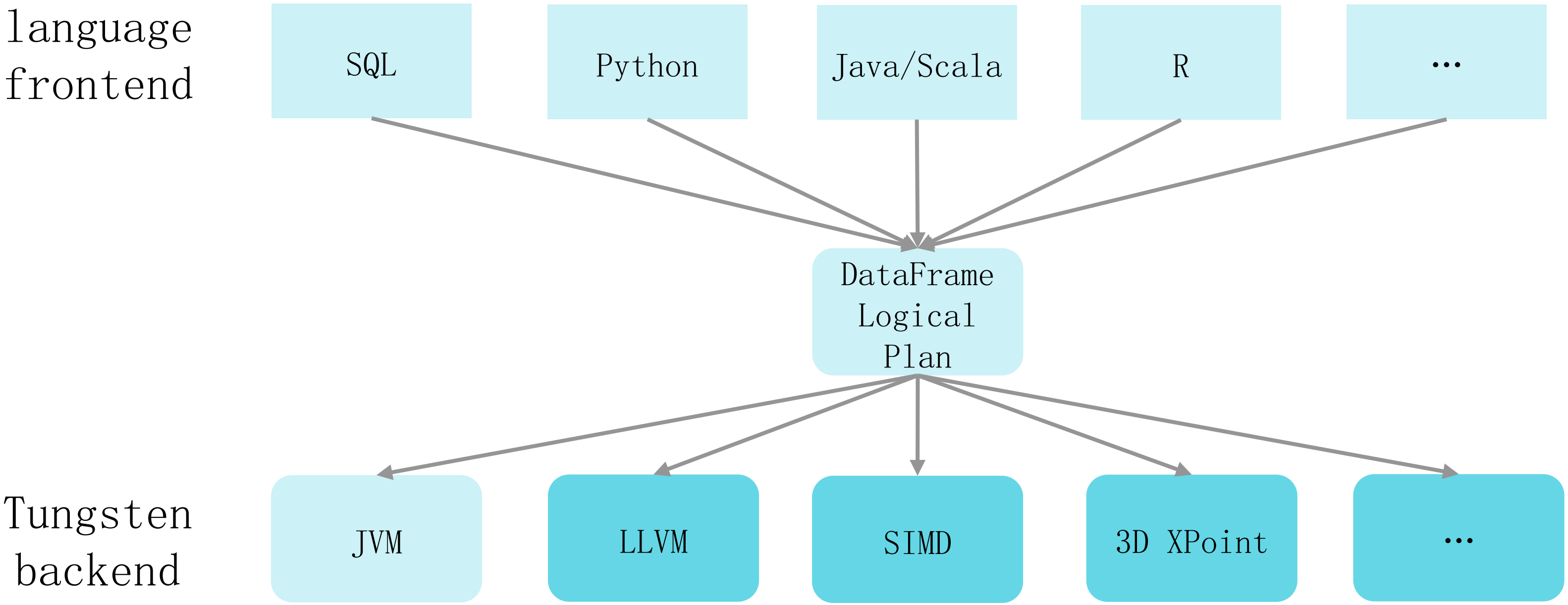
FOUNDERS
Jeff Bezos

3D XPoint

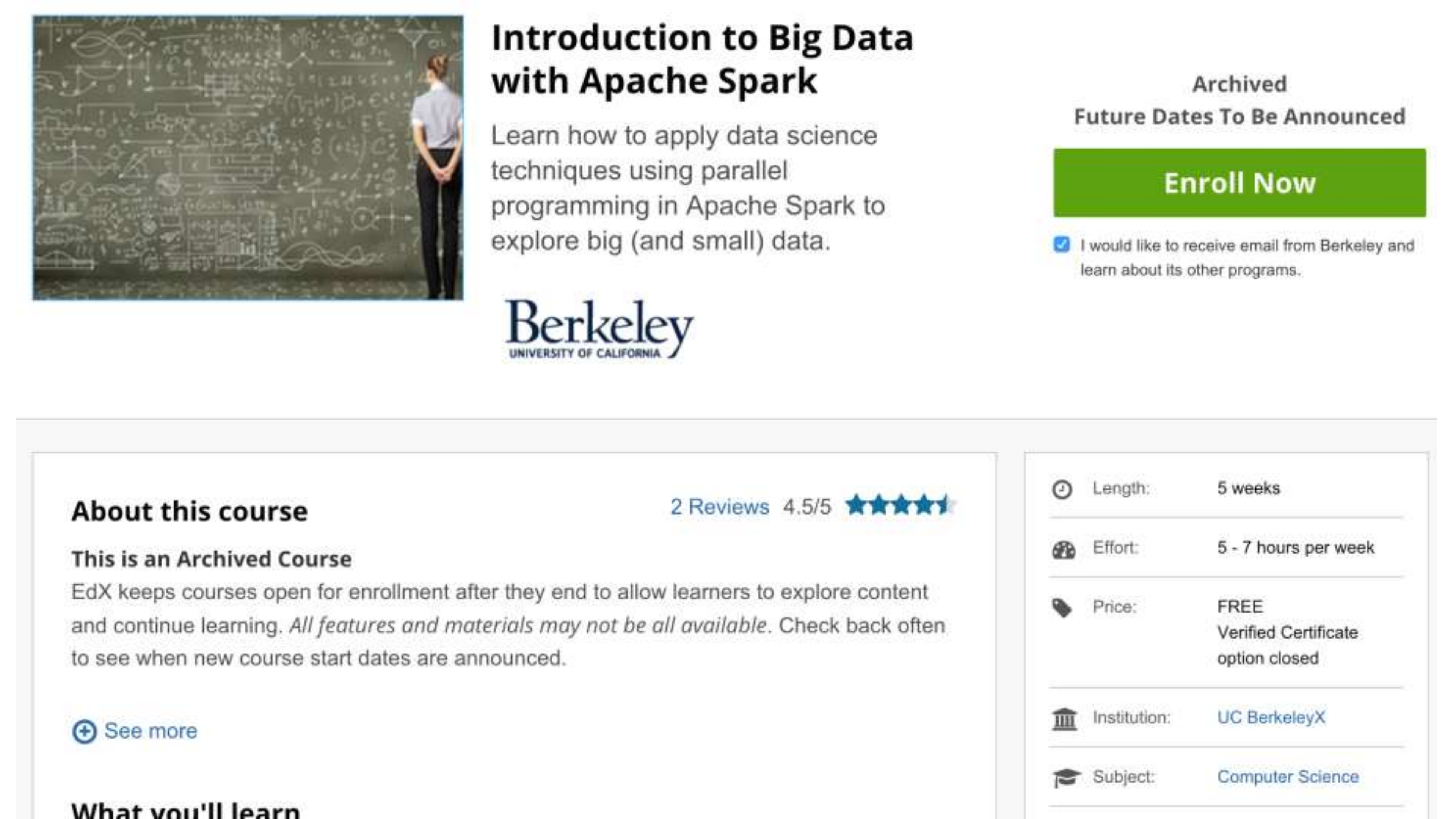


- DRAM latency
- SSD capacity
- Byte addressible

Unified API, One Engine, Automatically Optimized



databricks™



Introduction to Big Data with Apache Spark

Learn how to apply data science techniques using parallel programming in Apache Spark to explore big (and small) data.

Berkeley
UNIVERSITY OF CALIFORNIA

Archived
Future Dates To Be Announced

☑ I would like to receive email from Berkeley and learn about its other programs.

About this course

2 Reviews 4.5/5 ★★★★★

This is an Archived Course

EdX keeps courses open for enrollment after they end to allow learners to explore content and continue learning. *All features and materials may not be all available*. Check back often to see when new course start dates are announced.

Length:	5 weeks
Effort:	5 - 7 hours per week
Price:	FREE Verified Certificate option closed
Institution:	UC BerkeleyX
Subject:	Computer Science

What you'll learn

谢谢！

@rxin

databricks™