How Spark Usage is Evolving in 2015

Matei Zaharia

October 28, 2015

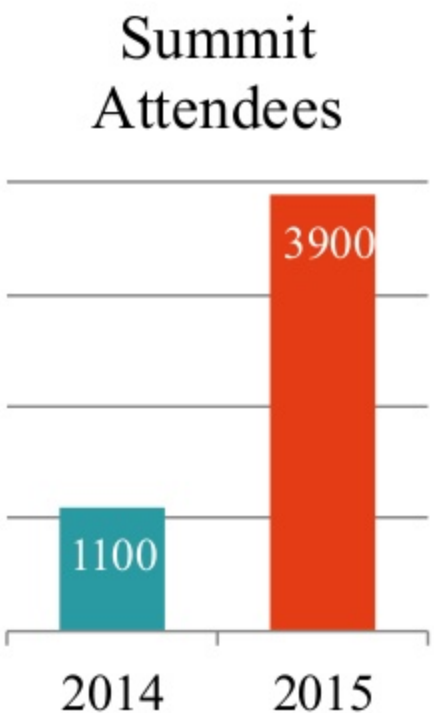
A Great Year for Spark

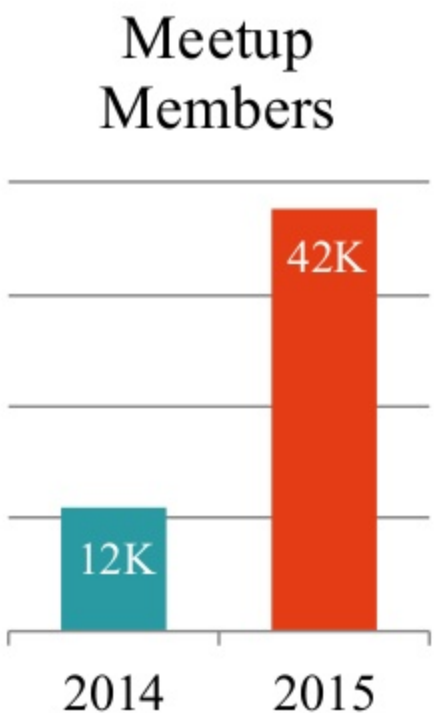
Most active open source project in big data

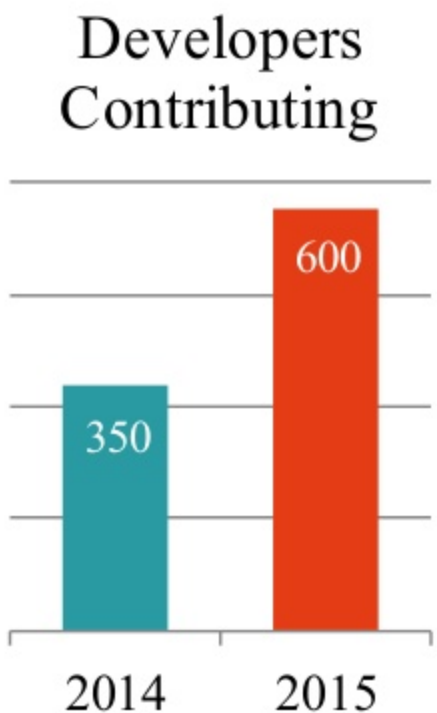
New language: R

Widespread industry support & adoption

Community Growth







Meetup Groups: January 2015

source: meetup.com

Meetup Groups: October 2015

source: meetup.com

What Spark Provides

General engine with libraries for many data analysis tasks

Access to diverse data sources

Simple, unified API

What Changed in 2015?

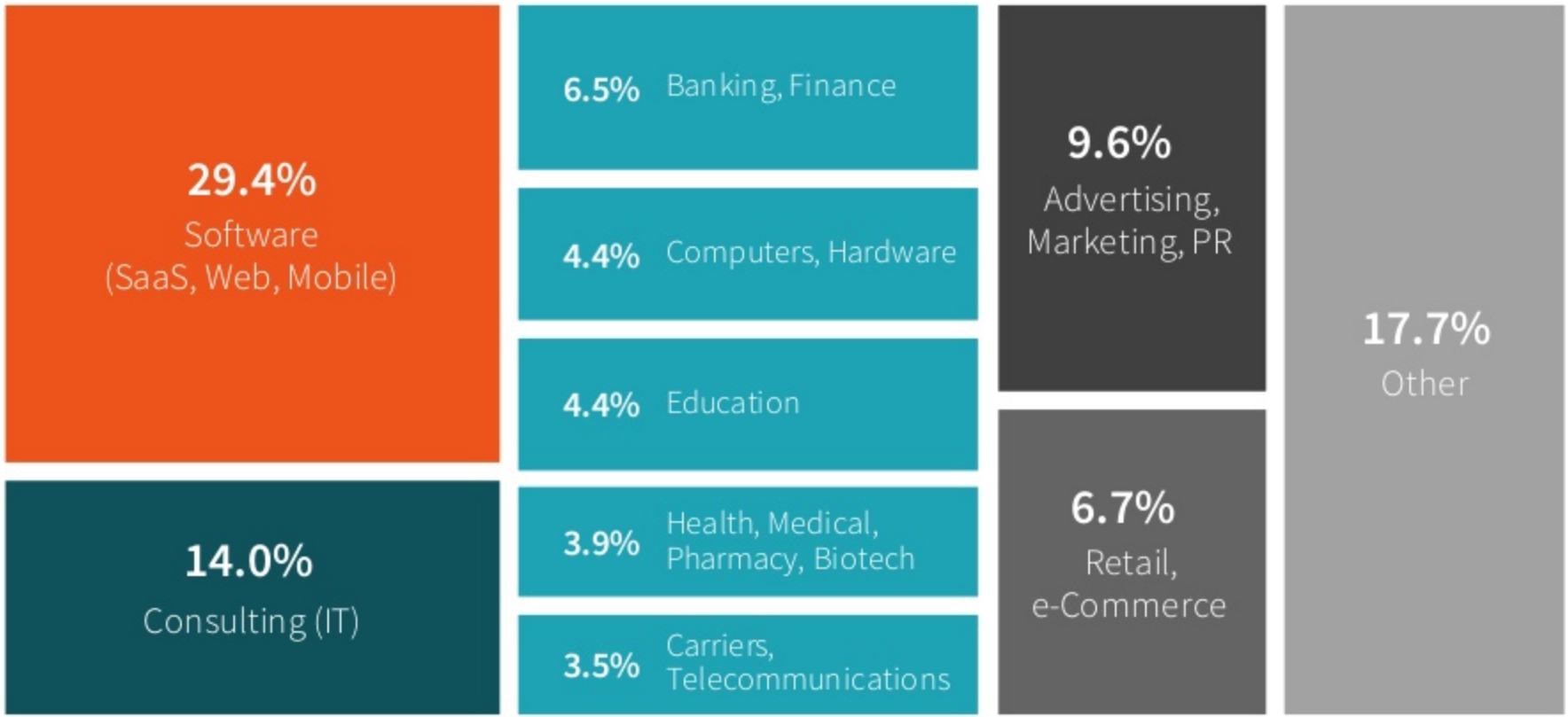
Databricks Survey

1400 respondents from 840 companies

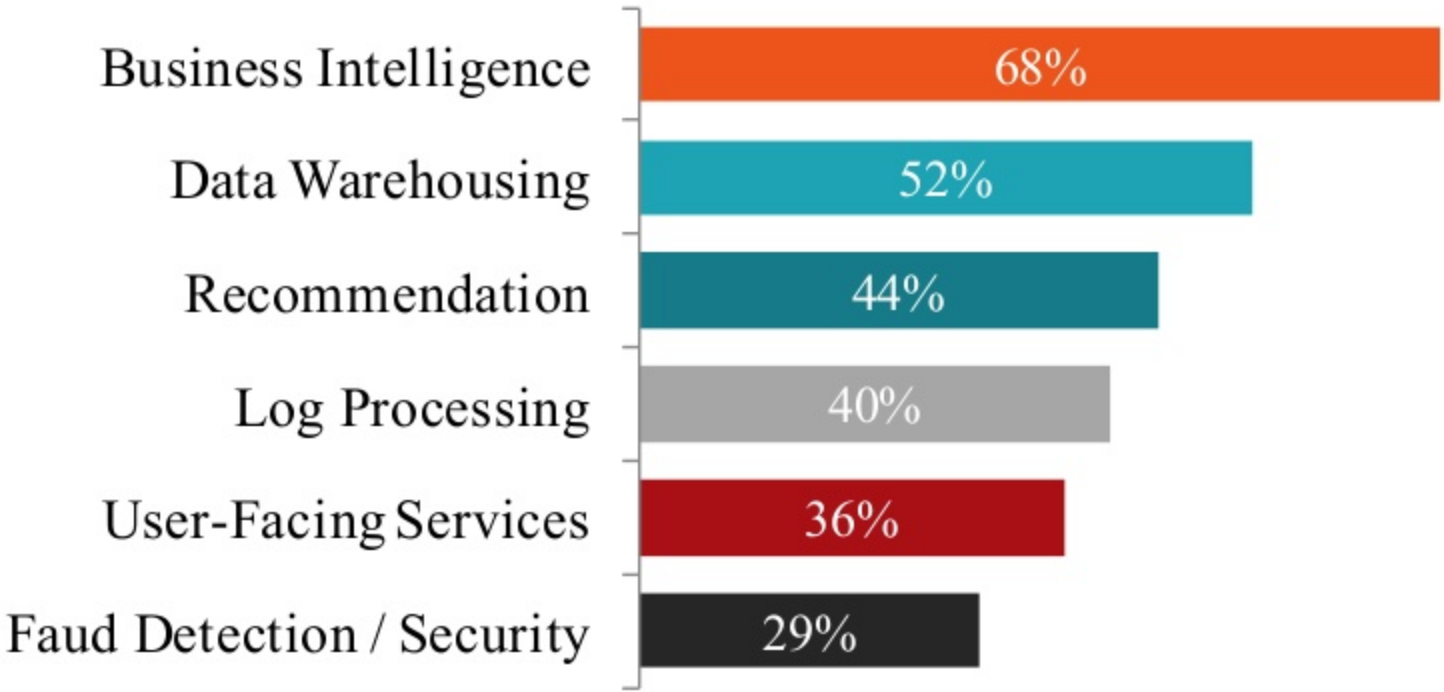
Three trends:

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

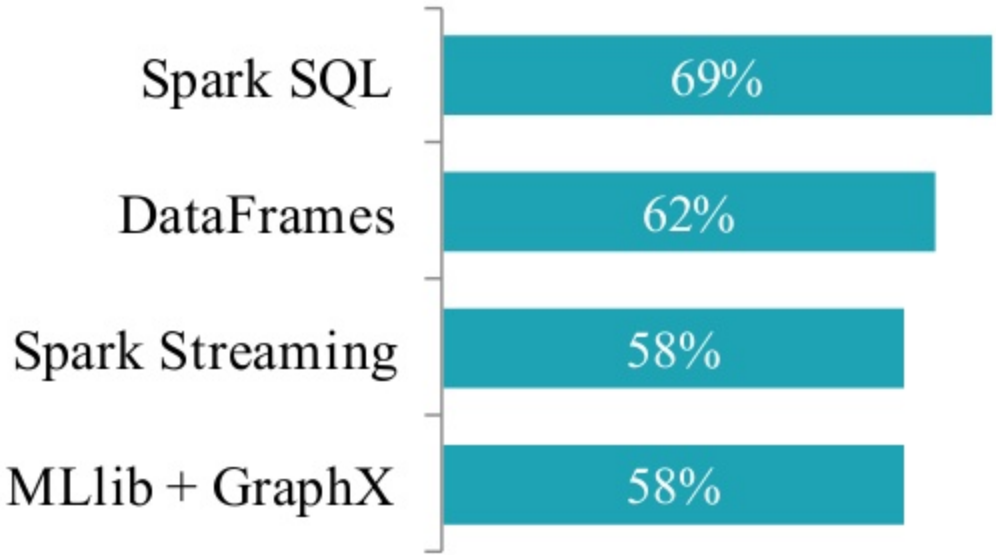
Industries Using Spark



Top Applications



Spark Components Used



75%

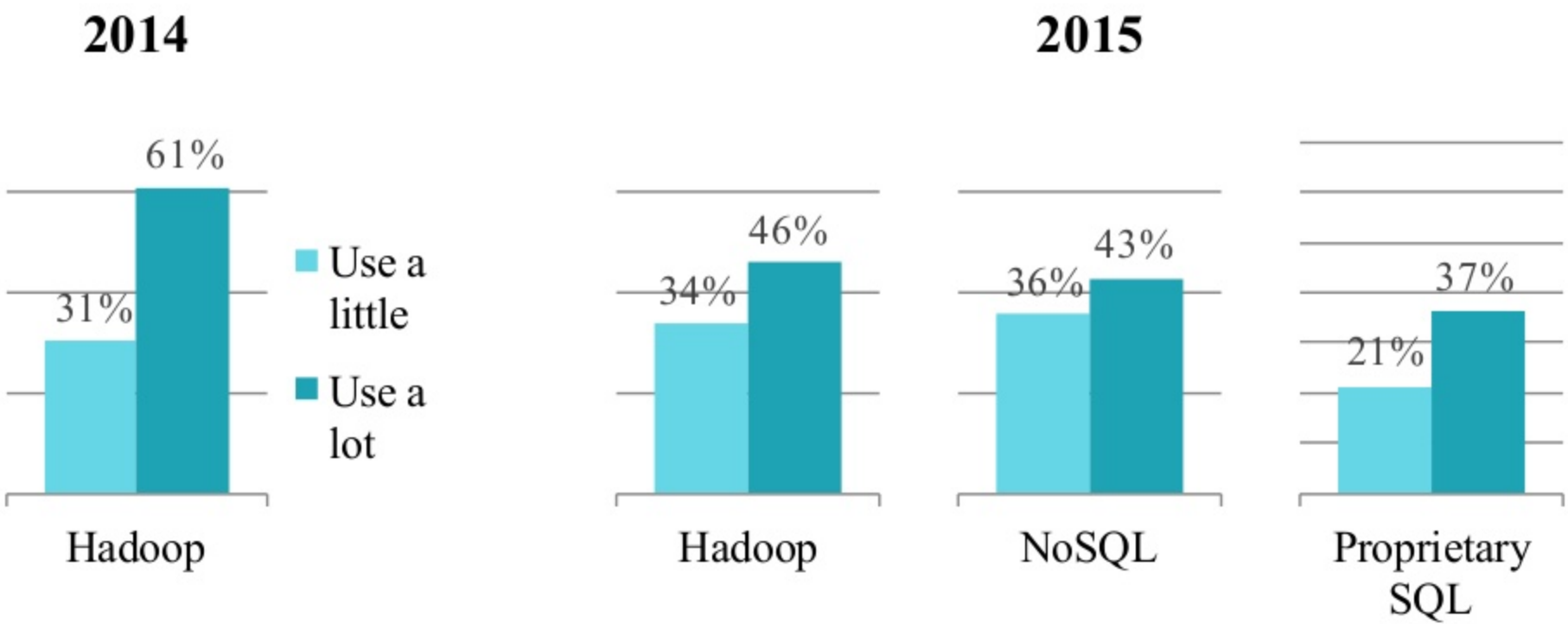
of users use more than one component

Diverse Runtime Environments

Hadoop: combined compute + storage

Spark: independent of storage layer

Diverse Runtime Environments



Diverse Runtime Environments

Cluster Managers

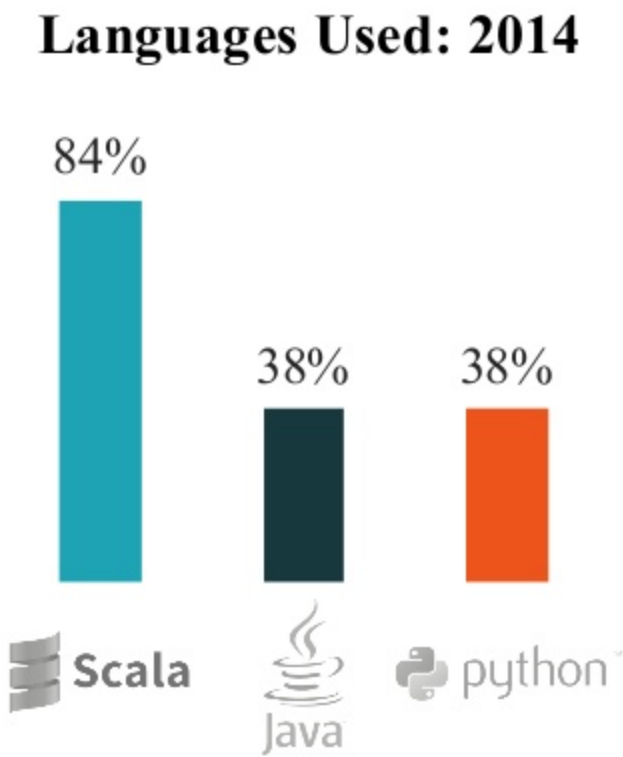
48% Standalone mode

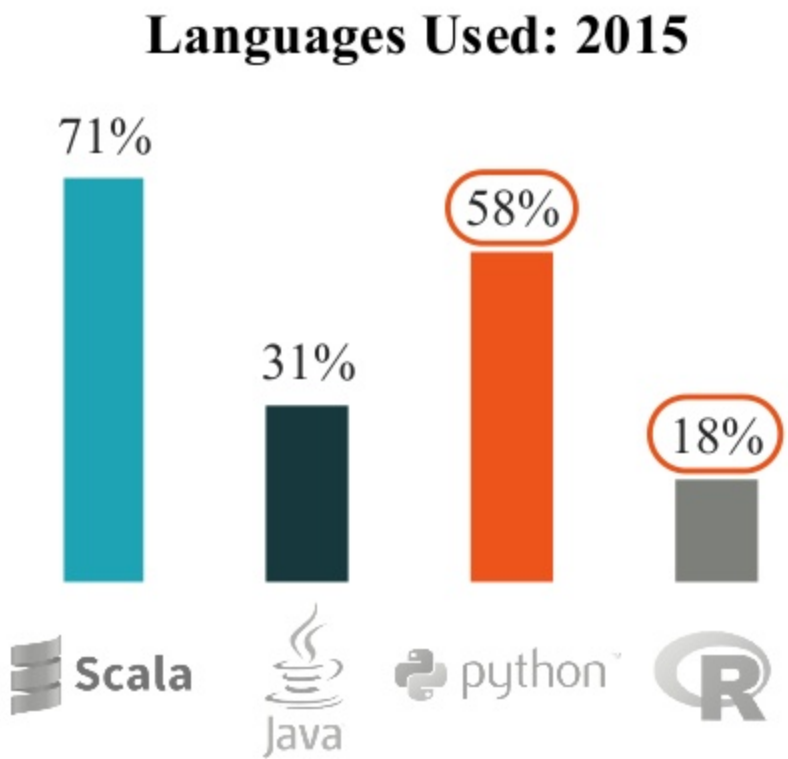
40% YARN

11% Mesos

51% on a public cloud

Diversity of Users





Fastest Growing Components

+280%

increase in
Windows users

+56%

production use
of Streaming

+380%

production
use of SQL

Are We Done?

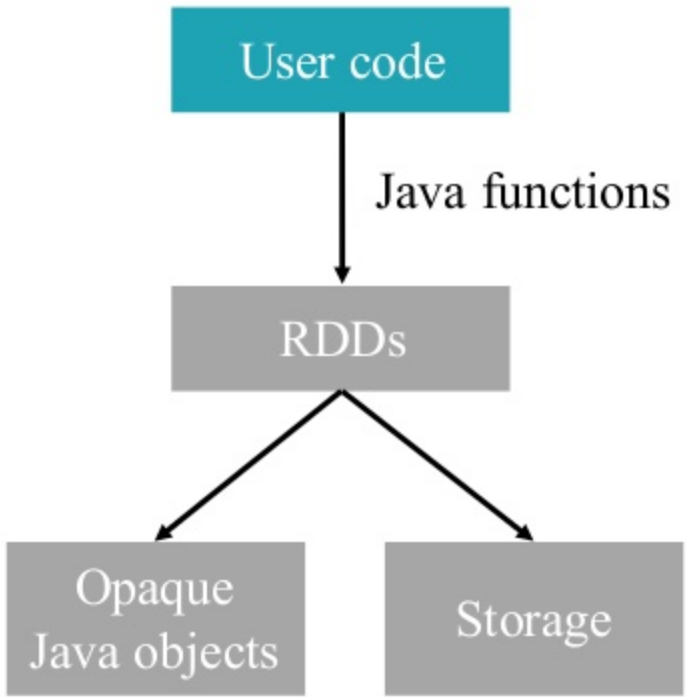
No! Development is faster than ever.

Biggest technical change in 2015 was DataFrames

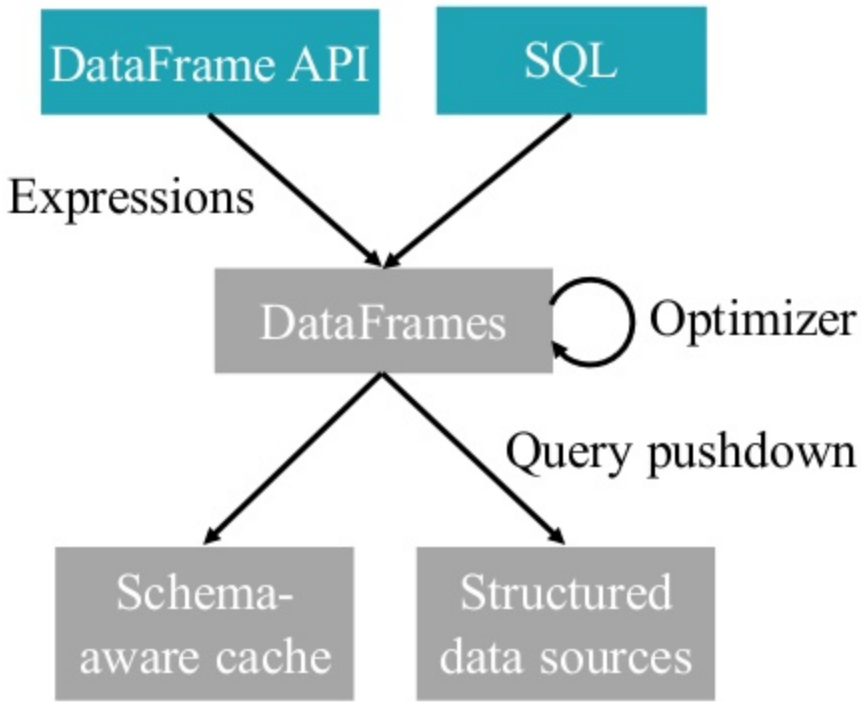
- Moves many computations onto the relational Spark SQL optimizer

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

Traditional Spark



DataFrames



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

See Reynold's talk tomorrow for details