



May 30, 2013

The Big Picture

Big data market is huge, and growing

» Hadoop is leading the charge

Our technology, Spark, outpaces Hadoop,
and gaining rapid traction

We are the best team for Spark

The Team



Ion Stoica (CEO)

- Professor and AMPLab co-director, UC Berkeley
- Conviva co-founder & CTO



Scott Shenker*

- Professor, UC Berkeley
- Nicira co-founder & CEO
- Most cited author in CS



Matei Zaharia (CTO)

- Dev. lead for Spark
- Mesos co-creator
- Hadoop committer (Fair Scheduler)



Ali Ghodsi, MBA

- Professor KTH, Sweden
- Researcher, UC Berkeley
- Peerialism co-founder (acq. by MPS)



Reynold Xin

- Dev. Lead for Shark
- Spark primary contr.



Andy Konwinski

- Mesos co-creator
- Google sched. team



Patrick Wendell

- Avro & Flume committer
- Mayfield fellowship



Arsalan Tavakoli

- Associate Principal, McKinsey
- PhD, Computer Science, UC Berkeley

*Chief Strategist

Big Data Analytics Today

Slow

- » Hadoop takes minutes even for simple queries

Limited

- » Fast systems (Impala, Redshift) don't go beyond SQL

Hard to use

- » On-premise clusters hard to set up, manage and scale
- » Cloud offerings still difficult to configure, and manage

Databricks: Next Generation of Data Analytics

Make simple things easy

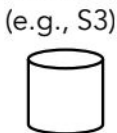
Make complex things possible

And everything fast

Make Simple Things Easy

Hosted service with interactive shell & dashboard

Cloud Storage
(e.g., S3)



5 mins



Interactive
Shell

15 mins



Dynamic
Dashboard



Business
Users



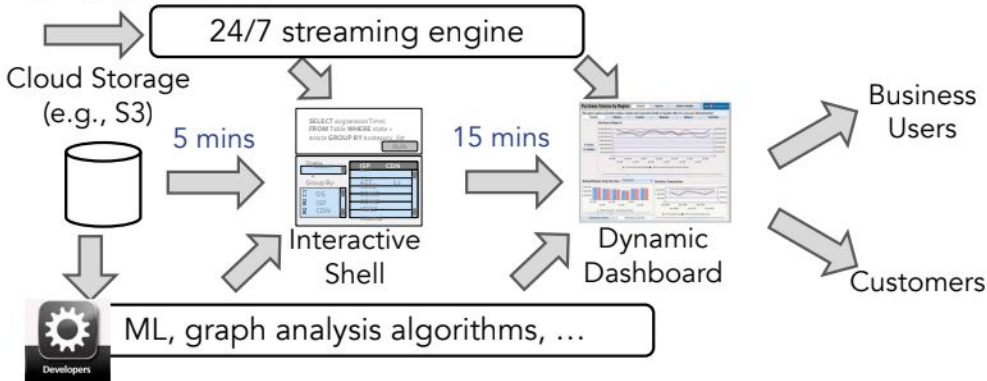
Customers

No cluster setup

No need to write code or understand algorithms

Just visit service and point at or upload data

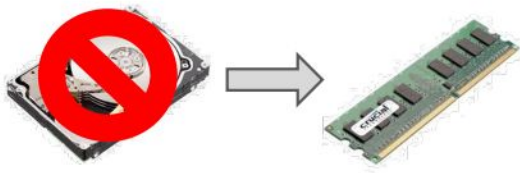
Make Complex Things Possible



General functionality:

- » Support all models: from SQL to ML
- » Support batch, interactive, and streaming
- » Enable developers to easily use/combine these models

And Everything Fast



Disks are the bottleneck

We avoid them as much as possible by aggressively using memory

Massive speed advantage over Hadoop

Core Technology: Spark

In development at Berkeley since 2009

Key innovation: Resilient Distributed Datasets
» Fault-tolerant, efficient in-memory storage

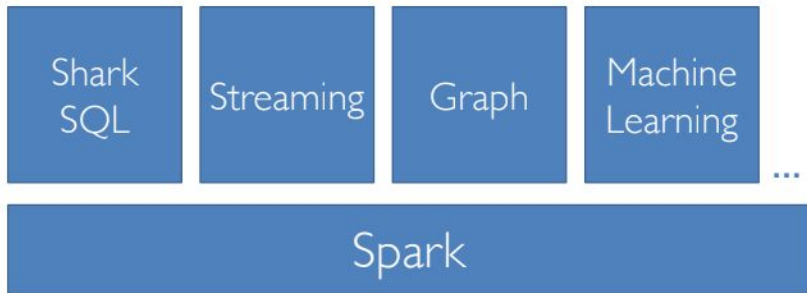
Fast: 100x faster than Hadoop

Easy to use: 5-10x less code than Hadoop

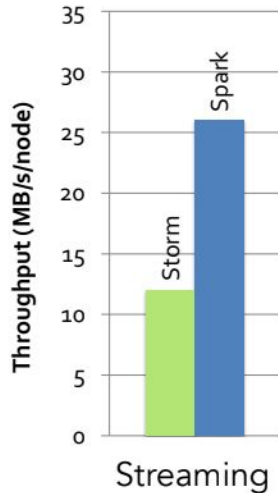
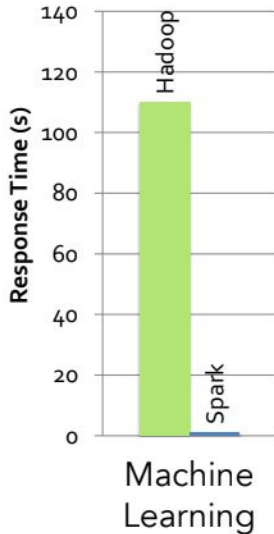
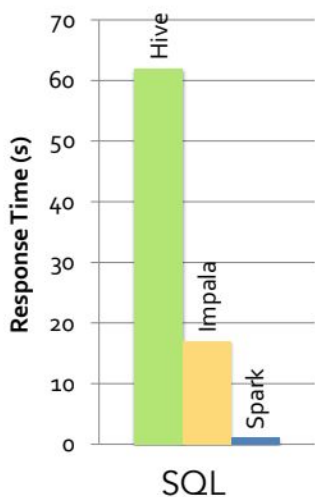
General

General Functionality

Superset of existing frameworks + ability to intermix at will



Performance & Generality



Gaining Rapid Traction



Sold out AMPCamp and
Strata tutorials

700+ meetup users

17 companies contributing
code

YAHOO!

intel

twitter

阿里巴巴
Alibaba.com

Adobe

webtrends

ClearStory
DATA
Now You See It

AdMobius

CONVIVA

bizo

TAGGED

quantifind

CONFIDENTIAL - DO NOT DISTRIBUTE!

Demo

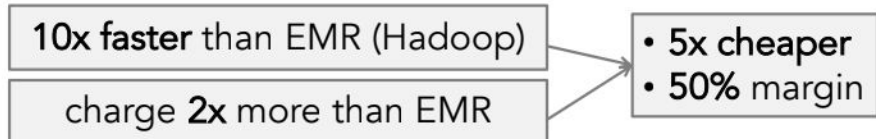
Initial Market Strategy

Start by offering cloud-based analytics platform

Easier for customers (no cluster management)

Easier for us (can control operating environment)

Speed advantage leads to reduced costs/prices



Next Steps

If some customers want on-premise

- » Virtual cluster appliance
- » Managed remotely
- » Still work from single code-base

If some customers want customization

- » We'll do it only if they write a big enough check!

Reach All Market Segments

Large and small organizations

Huge and small datasets

Experts and novices



First and last stop for data analytics,
and everything in between!

With Leading Technology

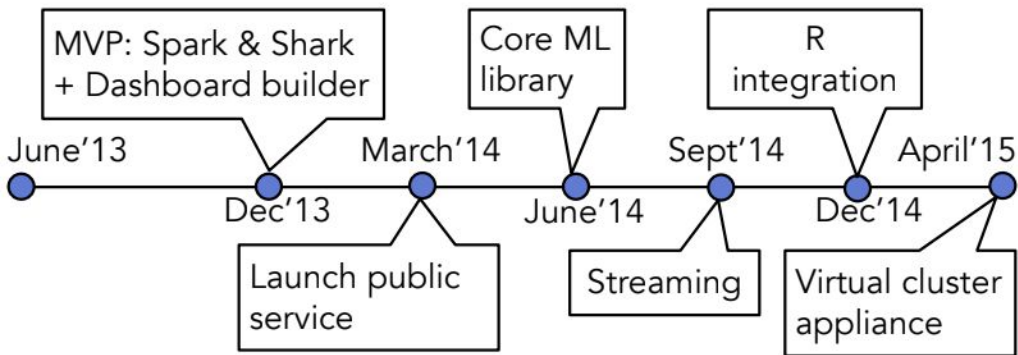
We will drive open-source Spark ecosystem

- » Our core technology will remain at cutting edge
- » Make Spark the de facto standard in academia

Will retain speed/cost advantages

- » Efficiency through multiplexing (leveraging Mesos)
- » Efficient internal management system

Product Roadmap



Projected Budget

	Q3'13	Q4'13	Q1'14	Q2'14	Q3'14	Q4'14	Q1'15	Q2'15
Employees	8	11	15	20	25	30	35	40
CapeEx & OpEx	\$440,000.00	\$615,000.00	\$887,500.00	\$1,275,000.00	\$1,600,000.00	\$1,900,000.00	\$2,225,000.00	\$2,505,000.00
COGS	\$0.00	\$0.00	\$0.00	\$240,000.00	\$700,000.00	\$1,350,000.00	\$2,400,000.00	\$2,500,000.00
Revenue	\$0.00	\$0.00	\$0.00	\$200,000.00	\$700,000.00	\$1,500,000.00	\$3,000,000.00	\$5,000,000.00
Profit	(440,000.00)	(615,000.00)	(887,500.00)	(1,315,000.00)	(1,600,000.00)	(1,750,000.00)	(1,625,000.00)	(5,000.00)
Balance	9,560,000.00	8,945,000.00	8,057,500.00	6,742,500.00	5,142,500.00	3,392,500.00	1,767,500.00	1,762,500.00

Assumption: \$10M series A

Goal: over 500 customers, and well over \$10M/year revenue by series B

A Simple Story

Ride a growing market: **Big Data**

Win with better technology: **Spark**

#1 ecosystem through open-source

Backup Slides

What is the Market Size?

Amazon: 2.5M EMR clusters in 2012, growing exponentially

- » \$100 revenue/cluster = \$250M
- » \$1,000 revenue/cluster = \$2.5B

AWS projected to reach \$20B revenue in 2020²

Microsoft Azure, \$1B revenue in 2012¹

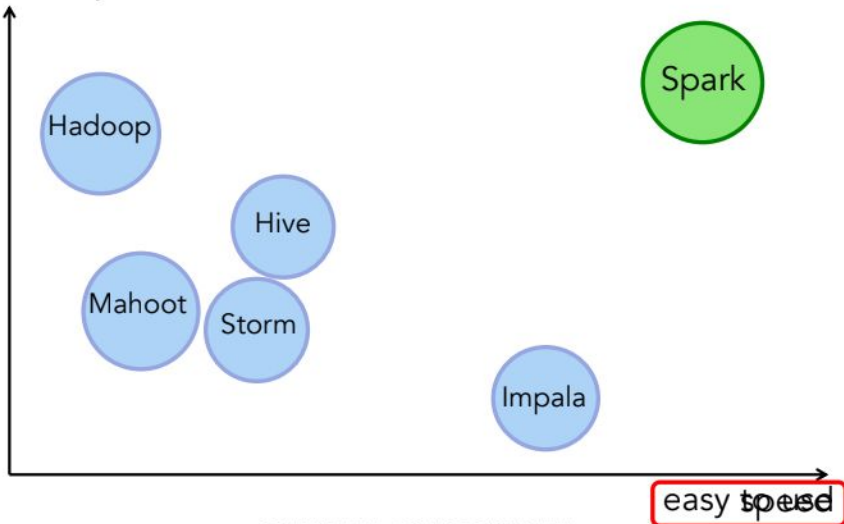
- » Expected to double annually

¹Blomberg

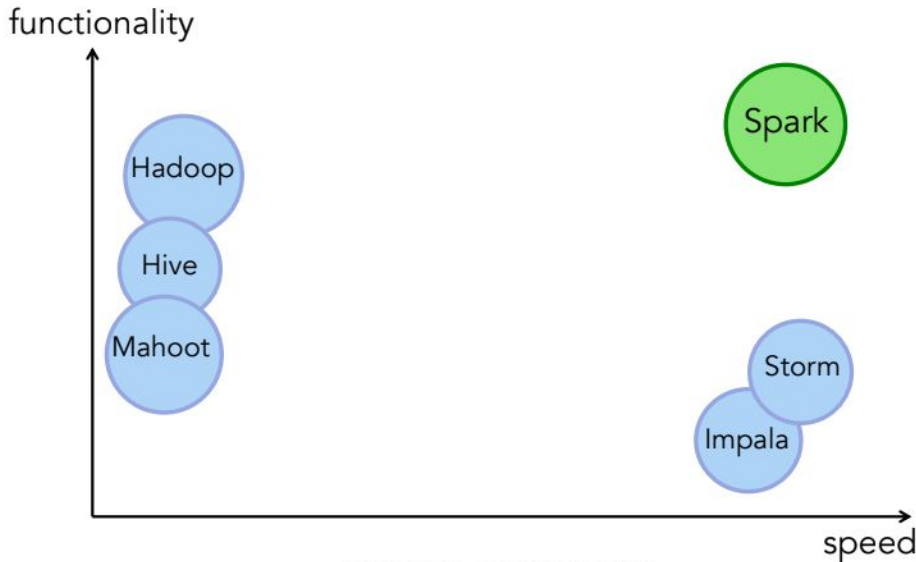
²Bernstein Research, 2013 report

Where does Spark Fit In?

functionality



Where does Spark Fit In?



Our Approach

Hosted platform based on Spark

- » Minimal set-up time and cost
- » Pay-as-you-go

GUI-based interactive data analytics

- » Interactive shell and dashboard builder



CONFIDENTIAL - DO NOT DISTRIBUTE!

How are we Different from BI?

UI not as powerful as existing BI tools

- » Cover only most popular features

Allow far more sophisticated computations

- » Graph based computations
- » Machine learning algorithms
- » ...

Similar to the difference between
calculator and first PC

Open Source Strategy

Successful open source → boost company

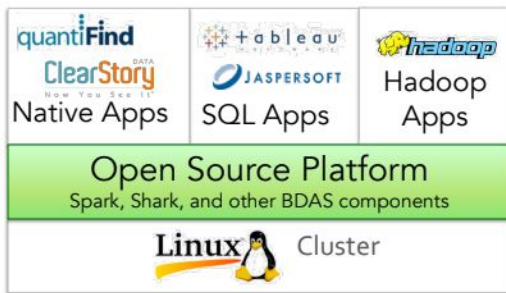
- » Leverage open source community to improve software
- » Enables a vibrant application developer community



Open Source Strategy

Successful open source → boost company

- » Leverage open source community to improve software
- » Enables a vibrant application developer community



Making Open Source Successful

Directly support two-three big companies

- » E.g., Yahoo!, Apple, Intel

Encourage Hadoop distributors to package, support, and provide professional service

- » E.g., MapR, Terradata

Train and certify developers and admins

- » DataBricks will stand behind them

Why Pipeline Builder and Dashboards?

Make simple things easy and complex things possible

Enable rapid growth

- » Eliminate dependency on 3rd parties to make platform accessible

Iron out APIs before exposing to 3rd parties

Encourage applications, even if better than ours

- » Recall, end game is to make platform dominant!

How do We Get There?

Lead open-source community

- » Harness energy of open-source developers



Training courses, meetups & tutorials



Educate next generation of big data scientists

- » Free use for data science courses

- » Make Spark platform of choice for big data education



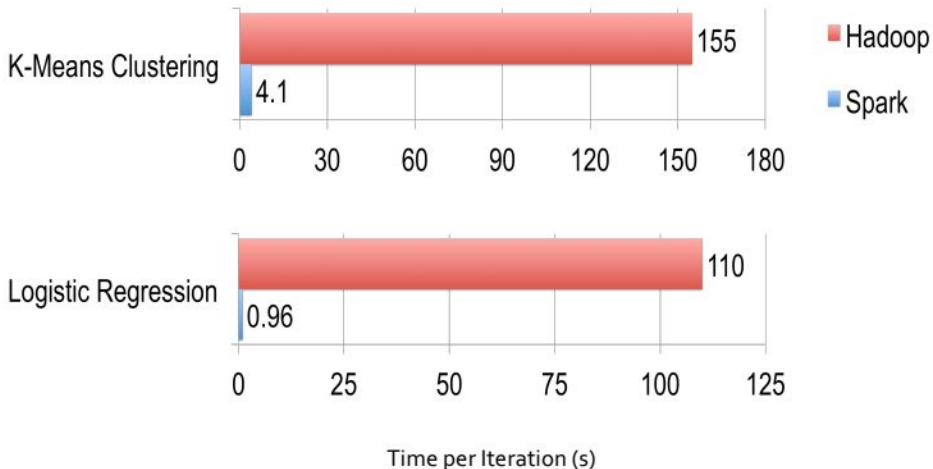
And Make Service “Sticky”

No reason for customers to leave

- » We are easier-to-use
- » We can meet all their functionality needs
- » We can keep price below what it takes to do it themselves

First and last stop for data analytics

Iterative Algorithms



An Analogy

1. Better phone, GPS, ...
2. New apps that combine features



First cellular phones



Specialized devices



Smartphone

An Analogy

Still limited and hard to use



Batch
processing

Specialized
systems

Unified
engine

Big Data Analytics Today

Primitive

- » Slow: Hadoop takes minutes even for simple queries
- » Limited: Impala, Redshift limited to SQL

Painful to use

- » Difficult and expensive to run on-premise cluster
- » Not cloud-friendly: hard to dynamically scale Hadoop