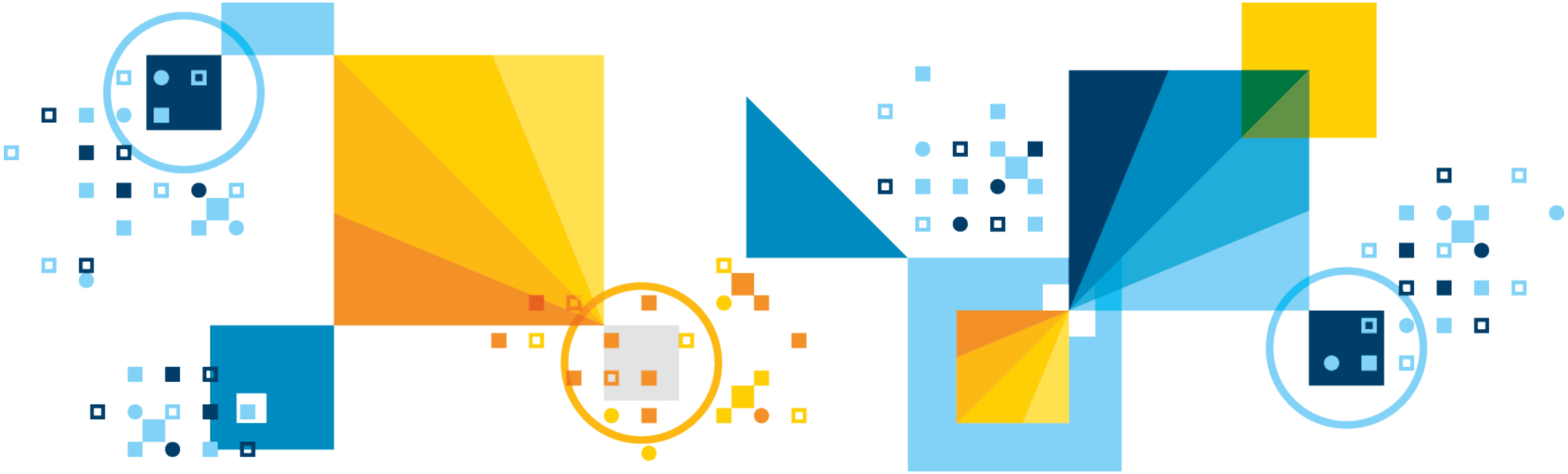
Djalma Cerino Filho
dcerino@br.ibm.com

16 Setembro 2015

Power of data. Simplicity of design. Speed of innovation.



Welcome from around the globe

Denver
Washington, D.C.
New York
Columbus
St. Louis
Chicago
Seattle
Minneapolis

Atlanta
Hartford
Dallas
Houston
Toronto
Dublin
London
Brussels

Moscow
Bonn
Paris
Milan
Helsinki
Stockholm
Oslo
Madrid

Tel Aviv
Warsaw
Melbourne
Singapore
Bangalore
Sydney
Kuala Lumpur

You are part of a global Spark community #SparkInsight

The insight economy is here

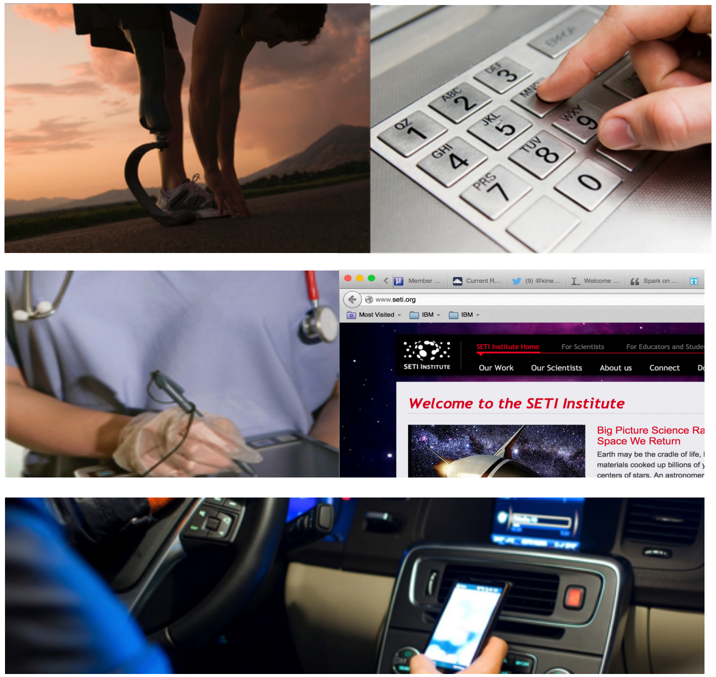
Front runners reap the benefits:

Analytics pay back $13.01 for every dollar spent[1]

69% created significant positive impact on business outcomes[2]

60% created significant positive impact on revenues[2]

53% created significant competitive advantage[2]



[1] Analytics Pays Back $13.01 for Every Dollar Spent" Nucleus Research, September 2014
[2] Analytics: The speed advantage" IBM Institute for Business Value, 2014

Imagine the possibilities

Real-time traffic flow optimization

Fraud and risk detection

Understand and act on customer sentiment

Accurate and timely threat detection

Predict and act on intent to purchase

Low-latency network analysis

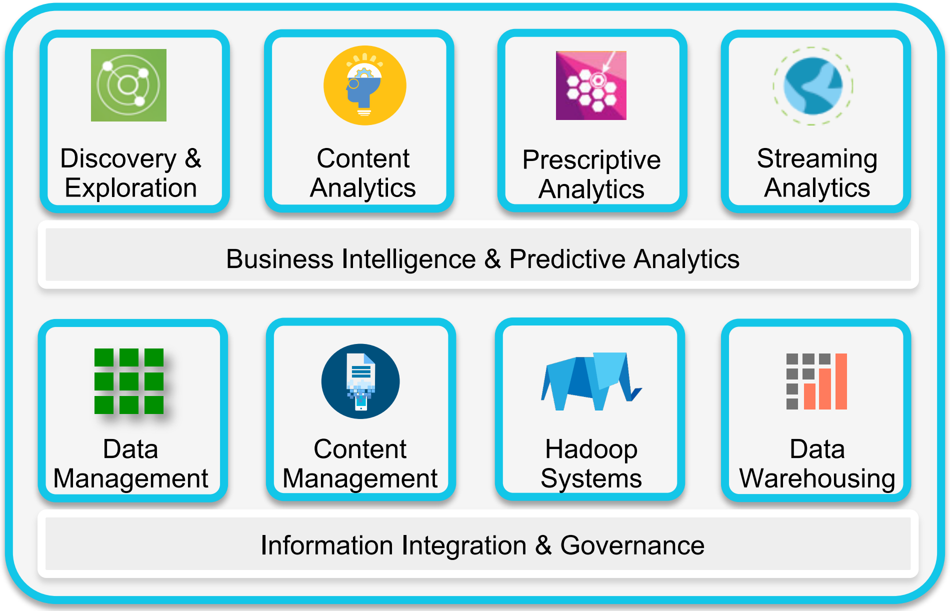
IBM Analytics Platform

Breadth and depth of analytics

Agile data integration and governance

Hybrid and fluid architecture

Open and unified platform



IBM announces major commitment to advance Apache® Spark™

…the most significant open source project of the next decade.

What is Spark?

An Apache Foundation open source project; not a product

An in-memory compute engine that works with data; not a data store

Enables highly iterative analysis on large volumes of data at scale

Unified environment for data scientists, developers and data engineers

Radically simplifies the process of developing intelligent apps fueled by data

Why Spark?

Spark is open, accelerating community innovation

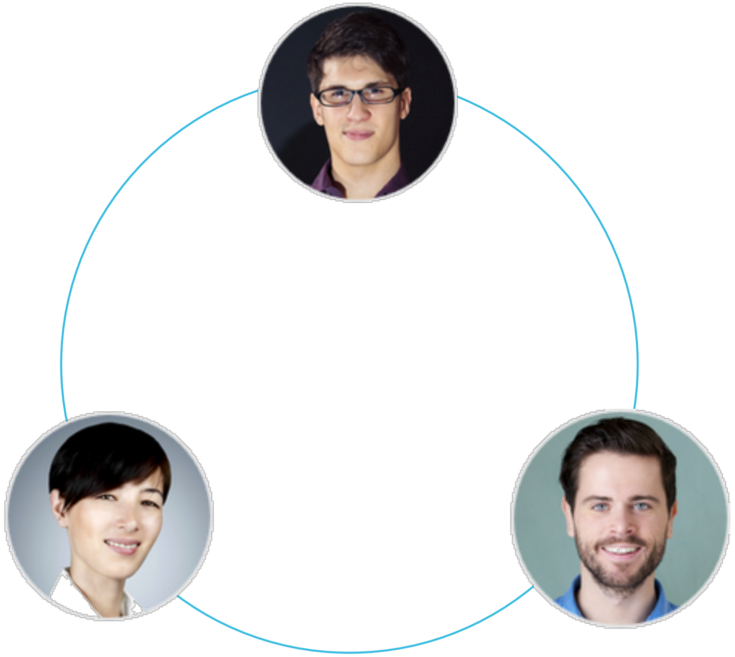
Spark is fast—100x faster than Hadoop MapReduce

Spark is about all data for large-scale data processing

Spark supports agile data science to iterate rapidly

Spark can be integrated with IBM solutions

Spark empowers users to accelerate the insight economy



Data Scientist

“the convincer”

Data Engineer

“the thinker”

App Developer

“the builder”

With Spark, data scientists can iterate models faster

Data Scientist

"the convincer"

What they want to do:

- Identify patterns, trends, risks, and opportunities in data
- Tell a story with data
- Discover new actionable insights
- Build new algorithms and models that move data science into the application

How Spark can help:

- Supports the entire data science workflow: from data access and integration, to machine learning, to visualization using the language of choice—typically Python
- Provides a growing library of machine learning algorithms via MLlib

With Spark, data engineers can build high-volume data systems

Data Engineer

"the thinker"

What they want to do:

- Bridge between the Data Scientist and the App Developer
- Implement machine learning algorithms at scale
- Put the right data system to work for the job at hand (Hadoop, Graph databases, Cloudant NoSQL, relational, streaming, in-memory)

How Spark can help:

- Abstract data access complexity (Spark doesn't care what your data store is)
- Enables near-real time solutions at web-scale (such as pipelined machine learning workflows)

With Spark, application developers can create analytics-driven apps

Application Developer

"the thinker"

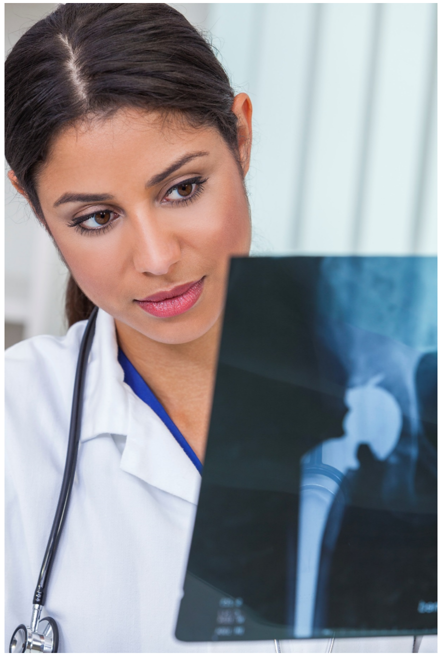
What they want to do:

- Build applications that lever advanced analytics in partnership with the data scientist and data engineer
- Follow agile design methodologies
- Optimize performance and meet SLAs

How Spark can help:

- Supports the top analytics app languages such as Python and Scala
- Eliminates programming complexity with libraries such as MLlib and simplifies DevOps
- Makes it easy to embed advanced analytics into applications

Clients have started innovating with IBM and Spark



Independence

NASA SETI INSTITUTE

NICE

the Analytics operating system

IBM | SPARK

The start of something big in data and design.

#SparkInsight

Our commitment to Spark

Announcing:

Open Source SystemML

Educate one million data professionals

Establish Spark Technology Center

Founding Member of AMPLab

Contributing to the core

Our largest contribution to open source since Linux

We are contributing SystemML

SystemML unifies the fractured machine learning environments

Gives the core Spark ecosystem a complete set of DML

Allows a data scientist to focus on the algorithm, not the implementation

Improves time to value for data science teams

Establish a de facto standard for reusable machine learning routines

Our investment to grow skills

Educate one million data scientists and engineers

Big Data University MOOC

- Spark Fundamentals I and II
- Advanced Spark Development series
- Foundational Methodology for Data Science

Partnerships with Databricks, AMPLab, DataCamp and MetiStream

Our goal is to be the #1 Spark contributor and adopter

Spark Technology Center

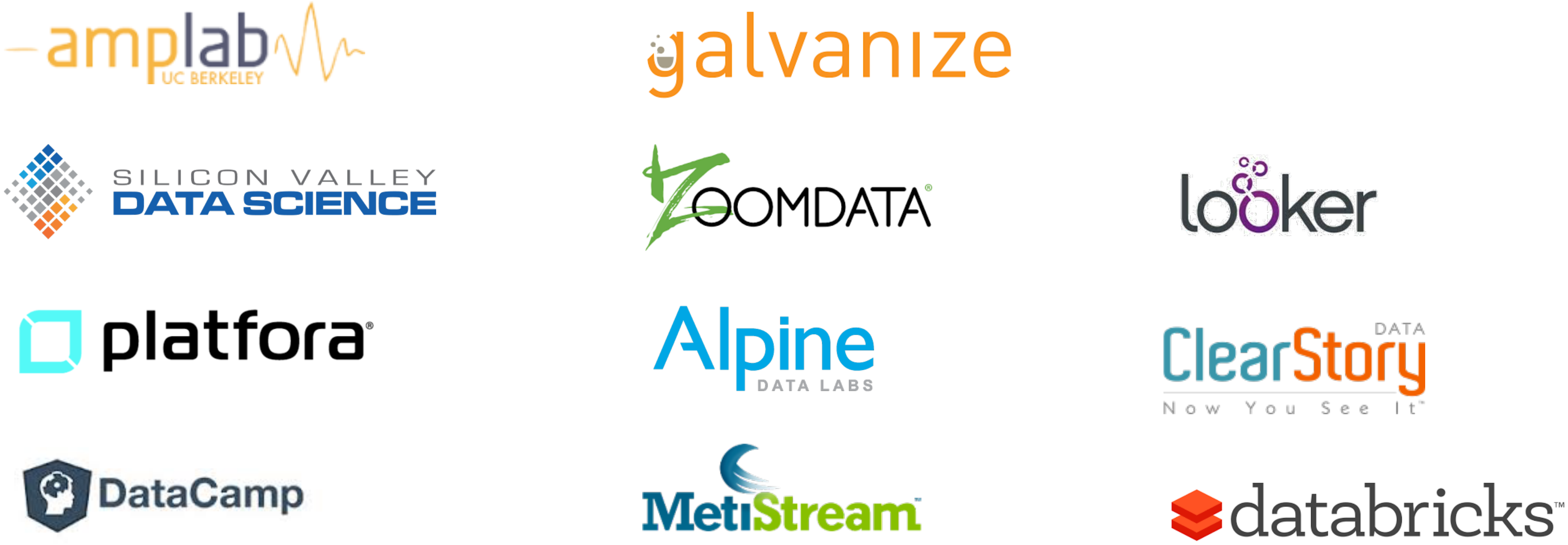
Inspire the use of Spark to solve business problems

Encourage adoption through open and free educational assets

Demonstrate real world solutions to identify opportunities

Use the learning to improve Spark and its application

Our partner ecosystem



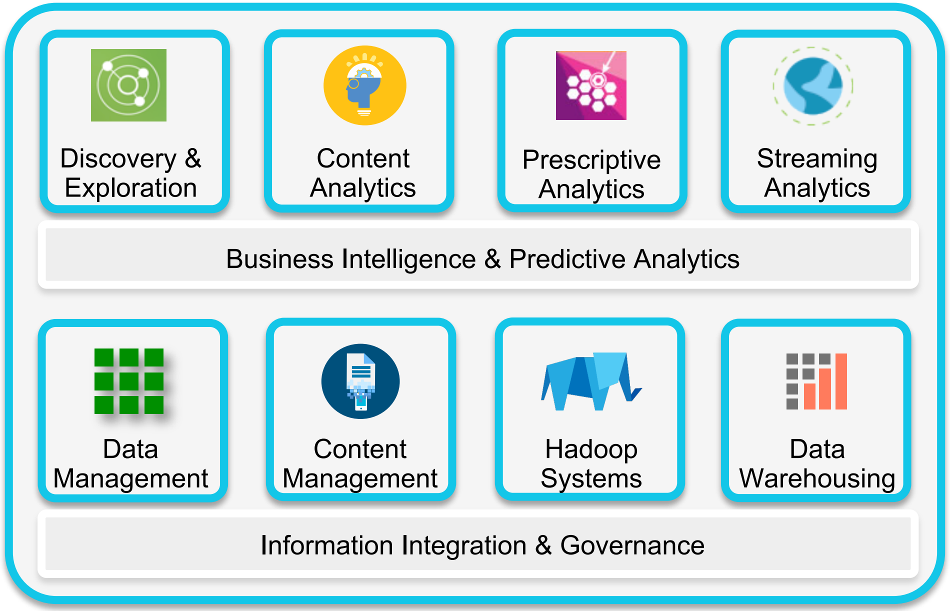
Spark is at work with our analytics platform

Spark

- Apache Spark as a Service on IBM® Bluemix™ (beta)

Hadoop Systems

- IBM Open Platform with Apache Hadoop can use Spark as alternative to MapReduce; supports all Apache Spark components
- IBM InfoSphere® BigInsights® modules intend to leverage Spark



Streaming Analytics

- Apply existing Spark models directly to IBM InfoSphere Streams
- Java Code written on Spark runs on IBM InfoSphere Streams
- Use same cluster for Spark and IBM InfoSphere Streams

Start with stampede to accelerate your outcome

1. Address common Spark use cases and intelligent applications
2. Domain-specific value in one day or two to three weeks
3. Knowledge transfer from IBM
4. Customize reference architecture and roadmap
5. IP that can be leveraged for business impact

Our use of Spark at IBM

Now

IBM Open Platform with Apache Hadoop

IBM InfoSphere Streams

IBM Platform Computing

Targeted for later in year

Apache Spark as a Service on IBM Bluemix (in beta)

IBM Watson™ Analytics

IBM SPSS® Modeler & Analytics Server

IBM DataWorks

IBM PureData™ Systems with Fluid Query

IBM Commerce

More than 30 IBM Research initiatives

100 incubated applications in 10 days

3,500 researchers and developers to Spark

Take your next step with IBM

Contact your IBM rep to schedule a deeper dive

Discover: Visit IBM Big Data Hub to read the latest news

Learn: Start with the “Spark Fundamentals” at Big Data University

Try Spark: Sign up for Apache Spark as a Service on IBM Bluemix at www.spark.tc/beta

Try Spark with Hadoop: Download at IBM.com/Hadoop

Engage: Join the IBM Spark Technology Center at www.spark.tc

Converse: #SparkInsight

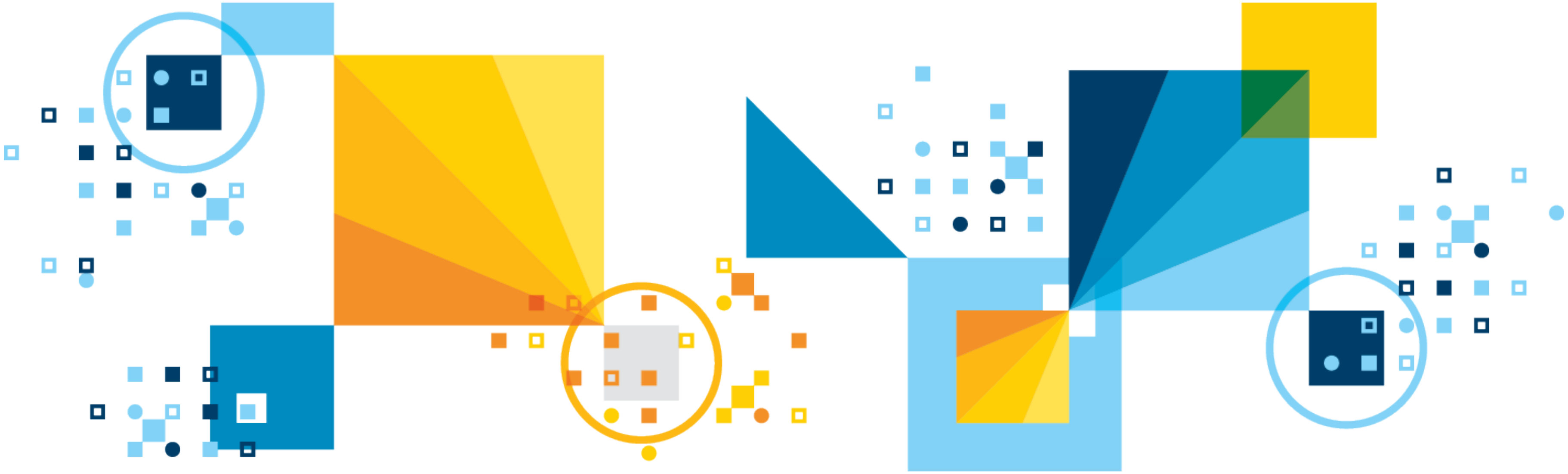
Why IBM?

Proven in analytics

Proven in open source

Proven in innovation

Power of data. Simplicity of design. Speed of innovation.



Additional Background

Apache Spark is an open source, in-memory compute engine that is fast, general purpose, and easy-to-use

Fast

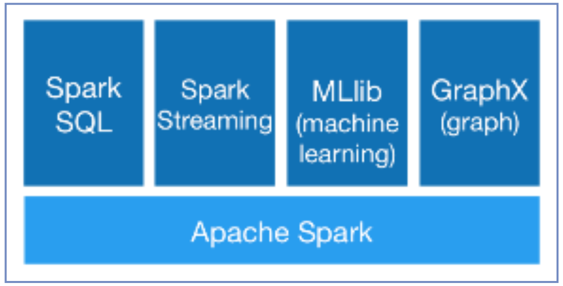
- Leverages aggressively cached in-memory distributed computing and JVM threads
- Faster than MapReduce for some workloads

General purpose

- Covers a wide range of workloads
- Provides SQL, streaming and complex analytics

Ease of use (for programmers)

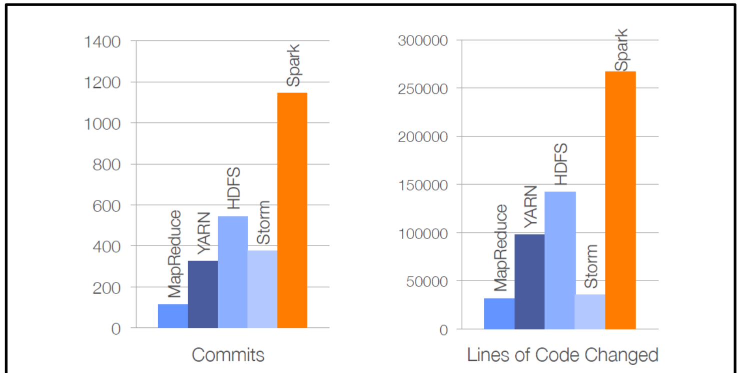
- Written in Scala, an object-oriented, functional programming language
- Scala, Python and Java APIs
- Runs on Hadoop, Mesos, standalone or cloud
- Scala and Python interactive shells



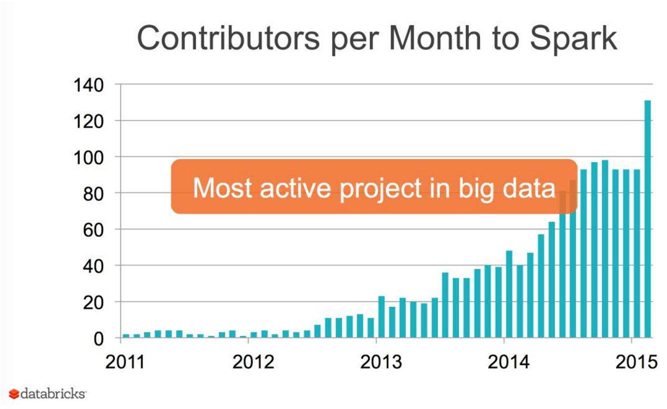
from http://spark.apache.org

Brief History of Spark

2002 – MapReduce @ Google
2004 – MapReduce paper
2006 – Hadoop @ Yahoo
2010 – Spark paper
2011 – Hadoop 1.0 GA
2014 – Apache Spark top-level
2014 – 1.2.0 release in December
2015 – 1.3.0 release in March
2015 – 1.4.0 release in June

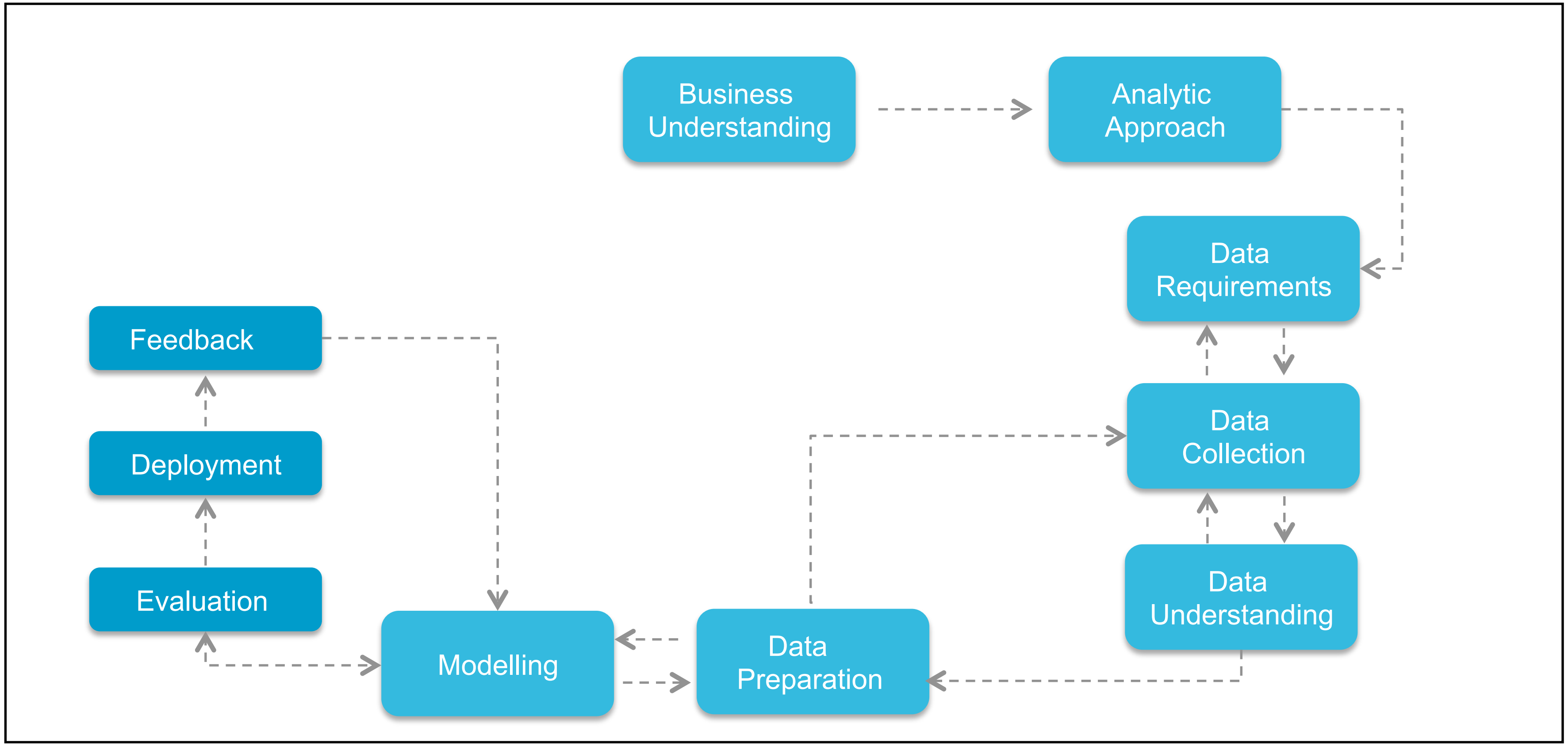


Activity for 6 months in 2014
(from Matei Zaharia – 2014 Spark Summit)



Spark is the most active project in Apache Software Foundation
Databricks founded by creators of Spark from UC Berkeley's AMPLab

Spark enables iterative cycle of data science



IBM Corporation
IBM Analytics
Route 100
Somers, NY 10589

Produced in the United States of America
July 2015

-

- This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

- THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

- Statements regarding IBM’s future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.