Survey Report

2015

databricks™

Introduction

01

Databricks ran our 2015 Spark Survey this summer to identify insights on how organizations are using Spark. The results reflect the answers and opinions of 1417 respondents representing 842 organizations.

Apache® Spark™ saw tremendous growth in 2014, and as the results of this survey demonstrate, Spark's growth comes not only from a huge increase in the number of contributors but also from increases in usage across a variety of organizations and functional roles.

Foreword

02

"Apache® Spark™'s core mission is to make Big Data simpler for enterprises of all sizes and across all industries, and the results of this year's Spark Survey validate this mission…"

"I'm excited to announce the results of this year's Spark Survey --not only because they indicate the rapid growth of the Spark community but also because they offer valuable insight into the direction Spark is moving. Spark's core mission is to make Big Data simpler for enterprises of all sizes and across all industries, and the results of this year's Spark Survey validate this mission. Today Spark is embraced by companies far beyond the IT industry, by a growing variety of functional roles within these companies (e.g. data scientists and analysts), and by users solving more complex data problems than ever before. And perhaps most exciting of all, Spark is growing well beyond Hadoop environments--a revelation that promises an exciting future for Spark."

MATEI ZAHARIA
CTO at Databricks,
VP Apache Spark at The Apache Software Foundation
@matei_zaharia

Spark Survey Report Highlights

03

TOP 3 APACHE SPARK TAKEAWAYS

Spark adoption is growing rapidly

Spark is growing far beyond Hadoop

Spark is increasing access to big data

SPARK IS THE MOST ACTIVE OPEN SOURCE PROJECT IN BIG DATA

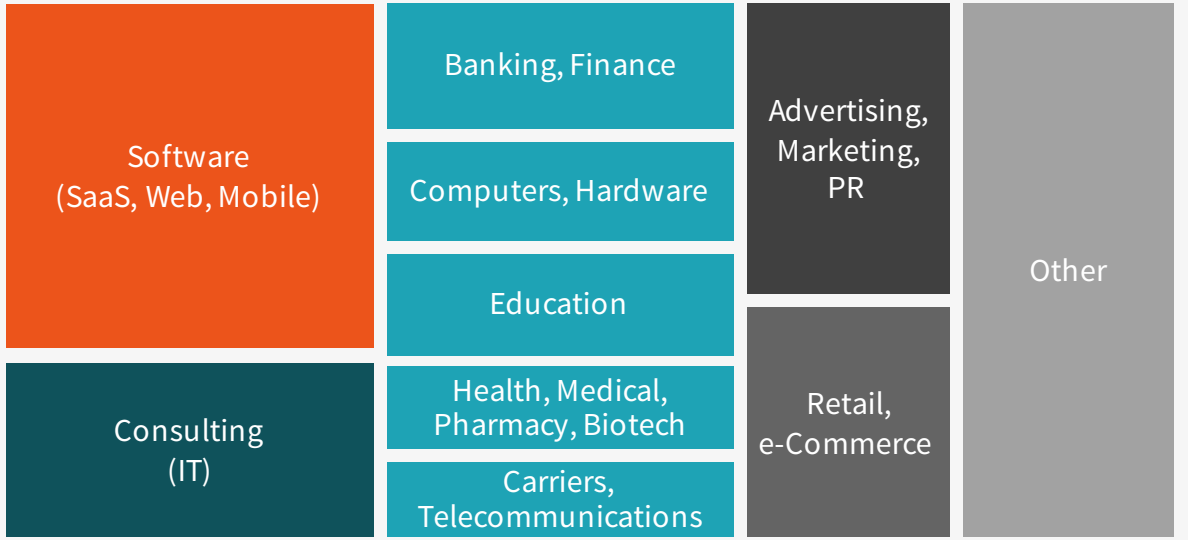
Spark Summit conferences

2014	2015*
1,164 attendees	2,986 attendees
453 companies	1,144 companies

Based on Spark Summit East and Spark Summit West, not including Spark Summit Europe

Spark contributors

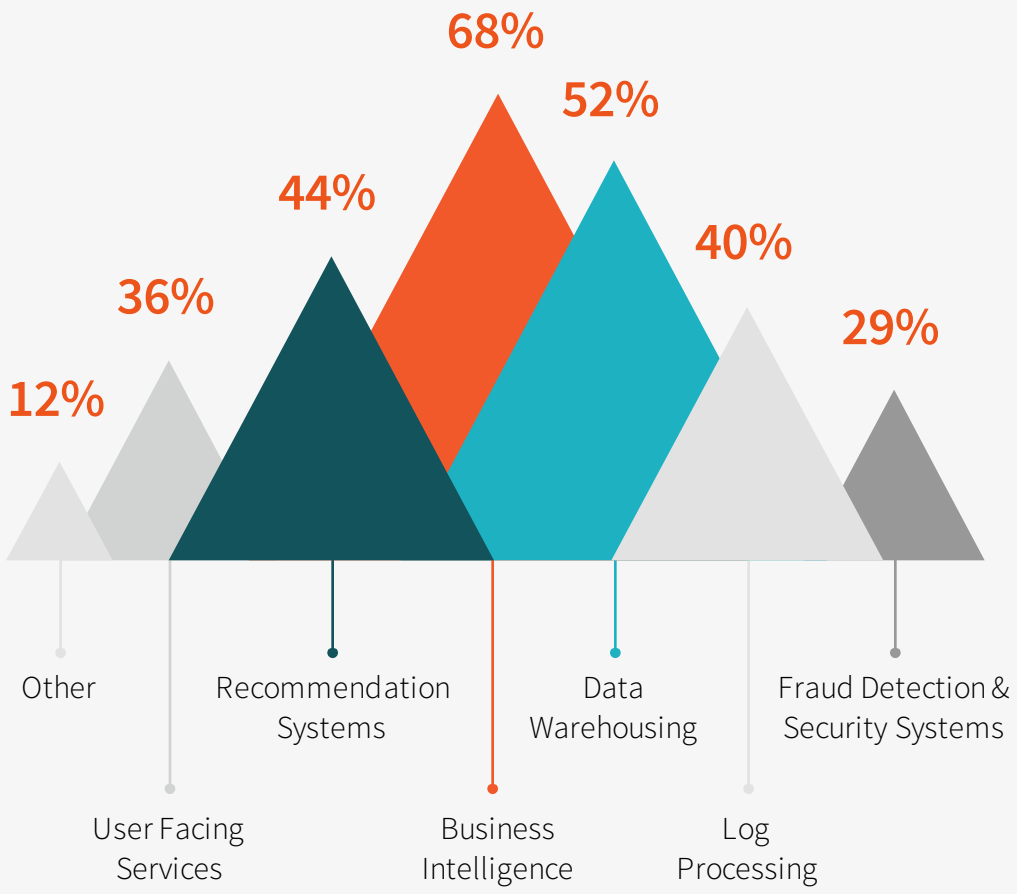
TOP 10 INDUSTRIES USING SPARK



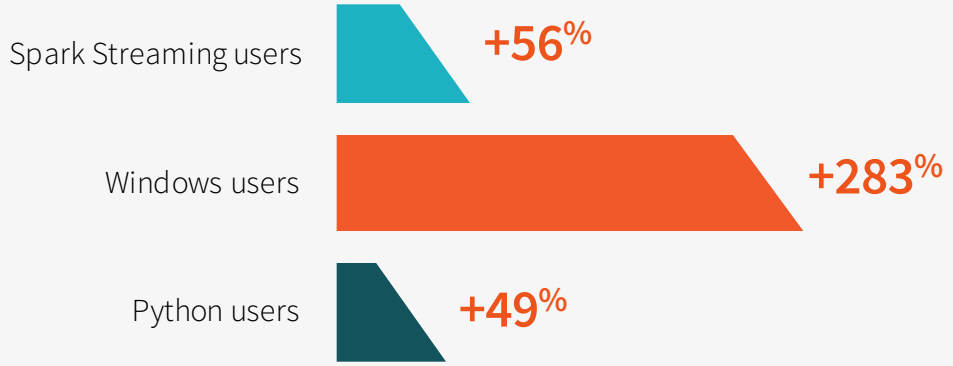
MOST IMPORTANT ASPECTS OF SPARK

91%	77%	71%	64%	52%
Performance	Ease of programming	Ease of deployment	Advanced analytics	Real-time streaming

SPARK IS USED TO CREATE MANY TYPES OF PRODUCTS INSIDE OF DIFFERENT ORGANIZATIONS



FASTEST GROWING AREAS FROM 2014 TO 2015



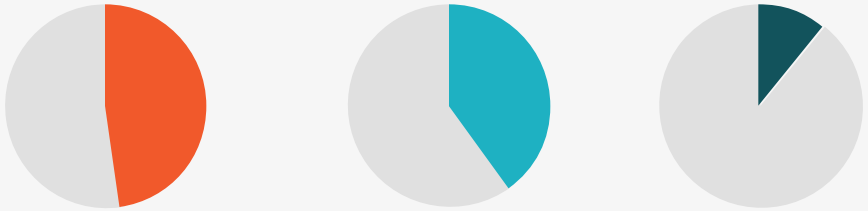
NOTABLE USERS THAT PRESENTED AT SPARK SUMMIT 2015 SAN FRANCISCO

Source: Slide 5 of Spark Community Update

NOVARTIS · UBER · AUTODESK · YAHOO! · salesforce · Microsoft · airbnb · Capital One · Alibaba.com · NBCUniversal · CISCO · Telefónica · Adobe · TOYOTA · Goldman Sachs · ebay · verizon · Palantir · NETFLIX · THOMSON REUTERS · NTT DATA

databricks

MOST COMMON SPARK DEPLOYMENT ENVIRONMENTS (CLUSTER MANAGERS)



48%
Standalone mode

40%
YARN

11%
Mesos

HOW RESPONDENTS ARE RUNNING SPARK

51%
on a public cloud

MOST USED SPARK COMPONENTS

69% | Spark SQL

62% | Dataframes

58% | MLlib + GraphX

48% | Streaming

75%

of Spark users are using two or more Spark components.

TOP ROLES USING SPARK

41%

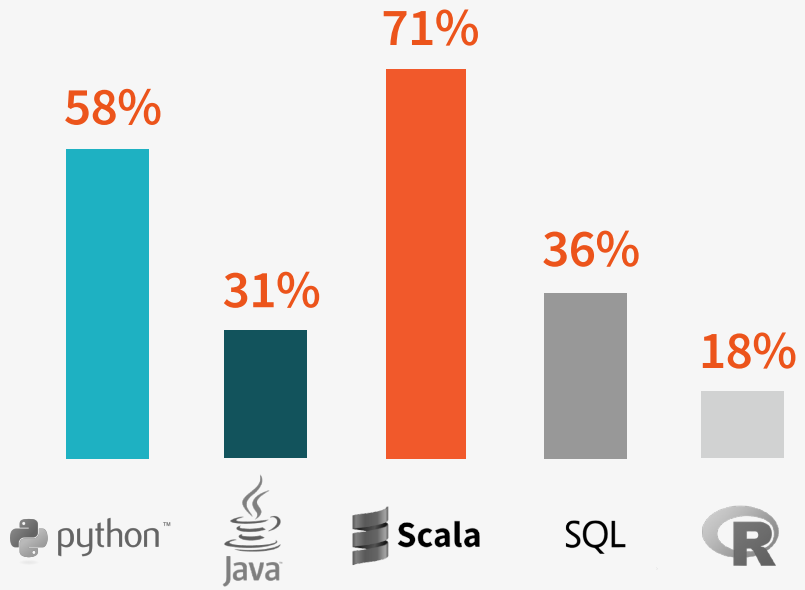
of respondents identify themselves as Data Engineers

22%

of respondents identify themselves as Data Scientists

PROGRAMMING LANGUAGES USED WITH SPARK

Survey respondents can choose multiple languages.



MOST IMPORTANT SPARK FEATURES

64% Advanced analytics

52% Real-time streaming

47% DataFrames

28% SQL Standards

databricks

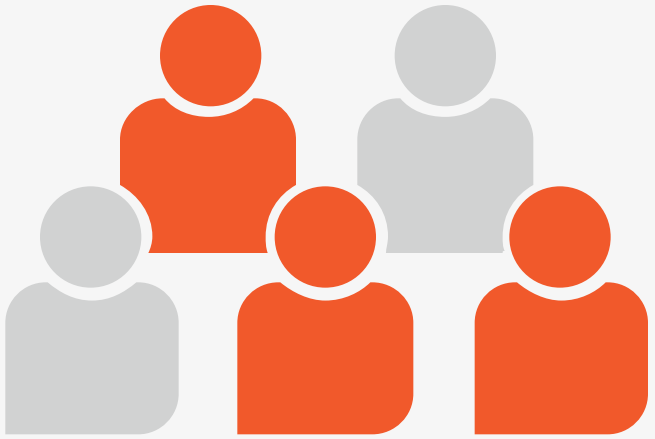
Spark Adoption is Growing Rapidly

04

SPARK PLATFORM IS GROWING

With over 600 Spark contributors in the last 12 months (315 last 12-24 months) Spark is the most active Apache Open Source project in Big Data.

More than 200 organizations contribute code to Spark, which makes Spark one of the largest communities of engaged developers. Spark Summit, which is the biggest gathering of the Spark community, has grown from 1000 attendees in 2014 to 3000 attendees in 2015 (between San Francisco and New York). Spark's growth has been unstoppable in 2015, and The Spark Survey reveals insight into the user trends driving this growth.



APACHE SPARK IS EXPANDING TO NEW AND DIVERSE USERS

Spark is growing fast–and a lot of Spark's growth may be attributed to increasing diversity in Spark's user base.

Spark is quickly becoming the data processing platform that can be used by everyone–not just data engineers. The 2015 Spark Survey demonstrates increased Spark use by data scientists writing in Python, Windows users across both technical and business teams, and expert developers with real-time use cases, alongside expansion across new industries. Adoption of Spark has spread beyond the technology industry to help companies address a growing variety of data problems.

+49% increase in Python users
(went from 39% to 58% of users)

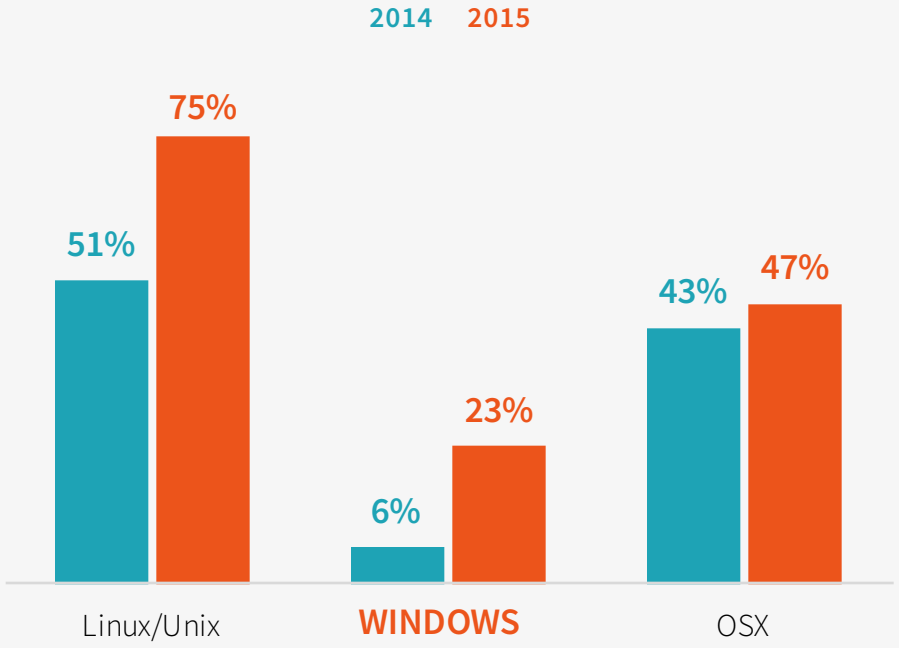
+283% increase in Windows users
(went from 6% to 23% of users)

+380% increase in SQL users
(went from 5% to 24% of users)

+56% increase in Streaming
(went from 9% to 14% of users)

FASTEST GROWING AREAS FROM 2014 TO 2015

ENVIRONMENT USED FOR SPARK DEVELOPMENT



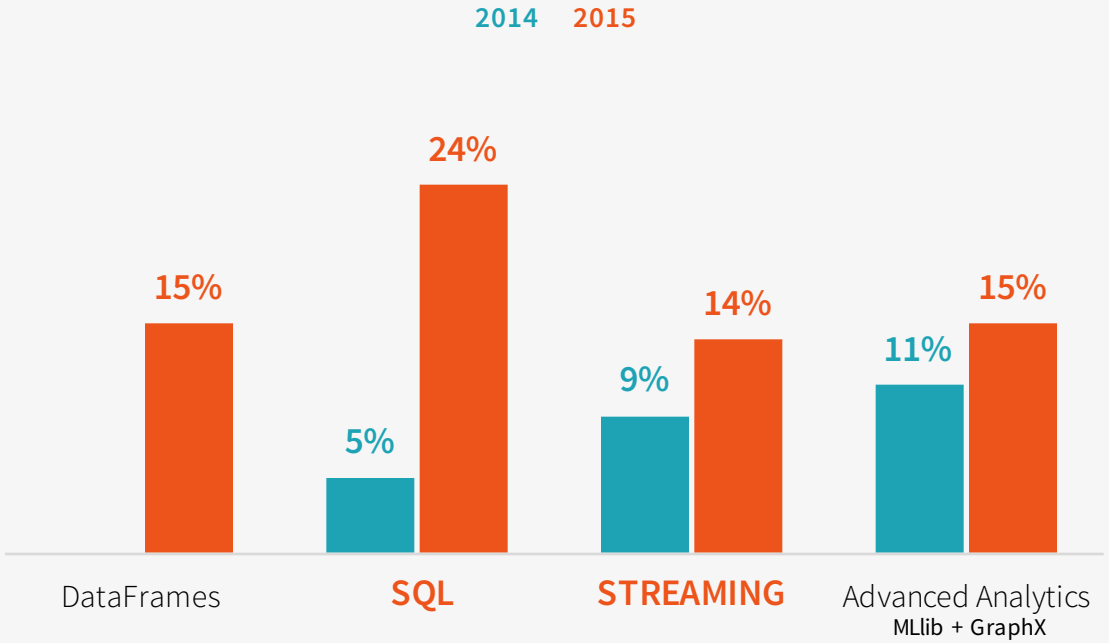
+283%

increase in Windows users

(went from 6% to 23% of users)

SPARK COMPONENTS USED IN PRODUCTION

(for all Spark Survey respondents)



+380%

increase in SQL users

(went from 5% to 24% of users)

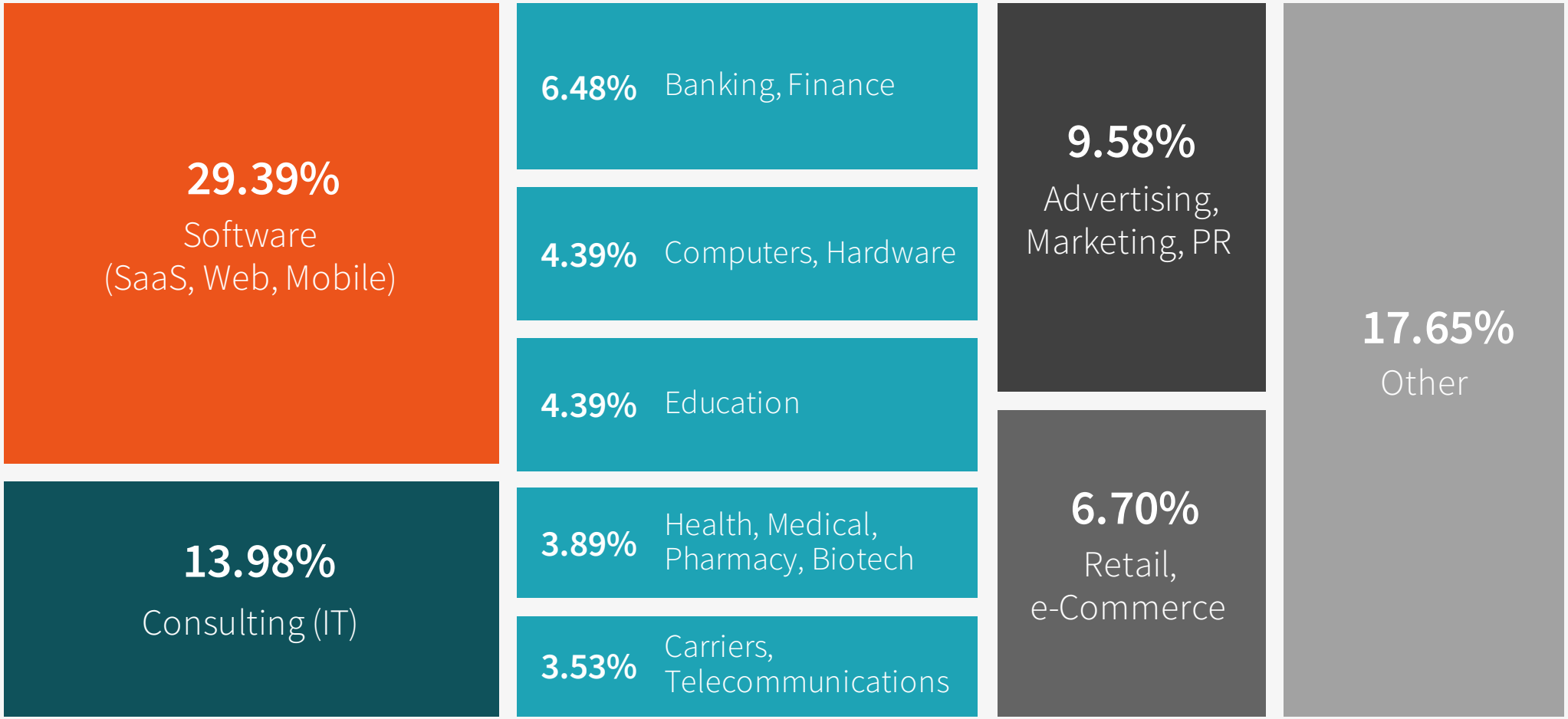
+56%

increase in Streaming

(went from 9% to 14% of users)

databricks

TOP 10 INDUSTRIES USING SPARK



USERS ARE GETTING STARTED RIGHT AWAY

Apache Spark's reliable performance and ease of use have driven much of the platform's expansion across new varieties of users. In 2014, Spark set the world record in large-scale sorting, a milestone that propelled Spark's performance into an industry standard. To maintain this position, Spark will have to demonstrate consistent performance improvements, the groundwork for which is already in place. Project Tungsten, the biggest performance improvement to Spark's execution engine since the project's inception, was launched in spring of 2015 to bring substantial improvements to the efficiency of memory and CPU for Spark's applications. As the Spark platform matures, it continues to push performance closer to the limits of modern hardware.

MOST IMPORTANT ASPECTS OF SPARK

91%	77%	71%	64%	52%
Performance	Ease of programming	Ease of deployment	Advanced analytics	Real-time streaming

MOST COMMONLY USED VERSION OF SPARK

64% of users run the most recent version

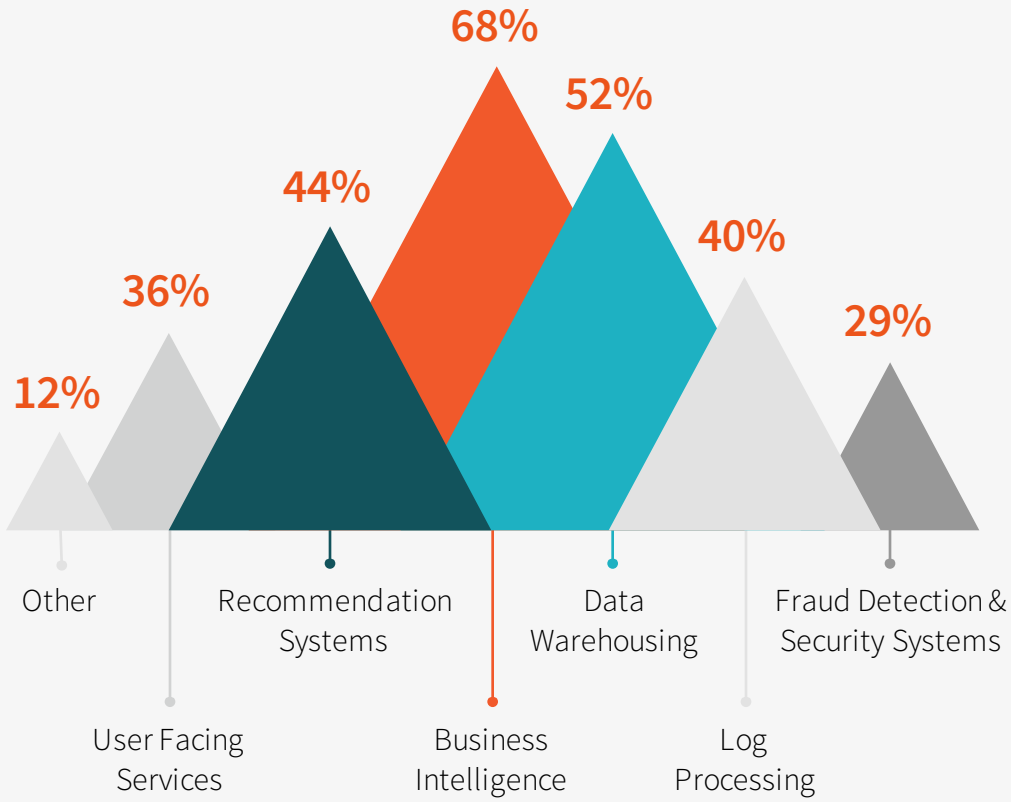
At the time of the survey, the majority of respondents were running Spark's latest update: Spark 1.4. In the fall of 2015, Spark released version 1.5, which includes under-the-hood improvements to performance, usability, and operational stability. The majority of Spark's users are fast to adapt new project updates–despite major user growth across new industries, new functional roles, and new data problems.

NOTABLE USERS THAT PRESENTED AT SPARK SUMMIT 2015 SAN FRANCISCO

Source: Slide 5 of Spark Community Update

USERS ARE SOLVING COMPLEX DATA PROBLEMS

Apache Spark is not only being used to solve an increasing variety of data problems but also an increasing complexity of data problems. Where traditional data processing vendors often prohibit companies from solving more than one data problem at a time, Spark enables users to build multiple data products simultaneously.



69% of respondents are creating two or more products with Spark

49% of respondents are creating three or more products with Spark

databricks

IN CLOSING...

Spark adoption is growing quickly as users find it easy to use, reliably fast, and aligned to growth in real-time & analytics.

Spark is Growing Far Beyond Hadoop

05

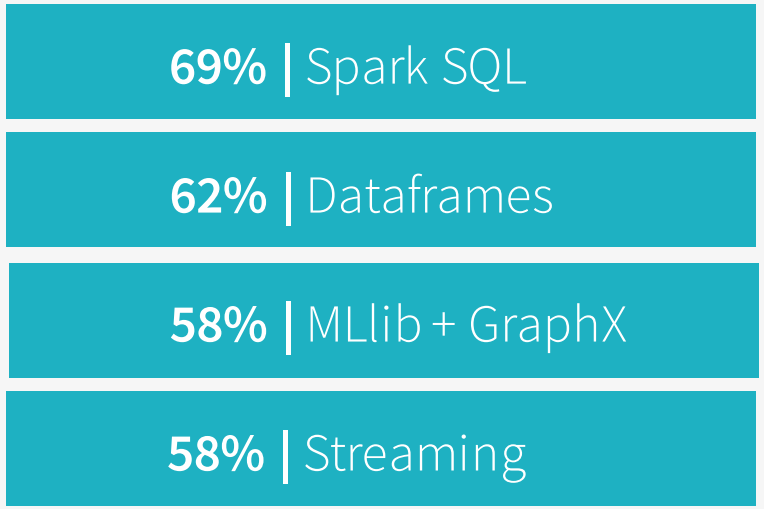
The rapid acceleration of Apache Spark adoption across new and diverse data problems is impressive. The fact that this growth is propelling Spark beyond Hadoop is astounding. Spark isn't a quick and easy add on to an existing data technology stack; Spark is the focus of a growing group of innovators that are driving tomorrow's data native culture.

MULTIPLE COMPONENTS OF SPARK ARE BEING USED IN THE SAME ORGANIZATION

Spark users aren't dividing and conquering each Apache Spark component separately; they're breaking down the barriers between data engineering and data science to work collaboratively across the Spark platform.

51% of Spark users are using three or more Spark components.

MOST USED SPARK COMPONENTS



75%

of Spark users are using two or more Spark components.

INCREASE IN PUBLIC CLOUD USE

Alongside growth of Apache Spark in standalone, use of Spark in the public cloud is now the primary deployment instance. Spark's users are opting to avoid complex deployment of the infrastructure in favor of instantaneous access on a public cloud.

HOW RESPONDENTS ARE RUNNING SPARK

51%
on a public cloud

INCREASE IN STANDALONE

While many users are running Spark within Hadoop and other data sources (e.g. Cassandra, HBase, etc.), a growing selection of users running Spark in standalone has surpassed those running Spark on YARN.

MOST COMMON SPARK DEPLOYMENT ENVIRONMENTS (CLUSTER MANAGERS)

48%
Standalone mode

40%
YARN

11%
Mesos

databricks

"Using Spark on 206 EC2 machines, we sorted 100 TB of data on disk in 23 minutes. In comparison, the previous world record set by Hadoop MapReduce used 2100 machines and took 72 minutes. This means that Spark sorted the same data 3X faster using 10X fewer machines."

Reynold Xin, Committer and PMC member of Apache Spark, Co-Founder Databricks

Spark is Increasing Access To Big Data

06

Apache Spark is creating opportunities for Big Data exploration by making it easier for wide range of people to solve a growing variety of data problems. It's not just distributed data engineers that want to work with Spark but also a growing constituent of data scientists.

INCREASE IN DATA SCIENCE USE

Apache Spark allows users to program in the language of their choice–an important capability that's driving Spark use across a growing audience of data scientists writing in Python and R. Where a 2014 Typesafe report identified 7.5 percent of Spark users as data scientists, 22 percent of this year's survey respondents identify as data scientists. As the presence of data scientists in the Spark community grows, so too does the popularity of programming languages Python and R. Support for Spark in R was released in June, 2015–weeks after The Spark Survey revealed that use of Spark in R is gaining momentum. Data scientists aren't just buying into Spark; they're diving into data problems right away.

TOP ROLES USING SPARK

41% identify themselves as Data Engineers

22.2% identify themselves as Data Scientists

17.2% identify themselves as Architects

10.6% identify themselves as Management

6.2% identify themselves as Academia

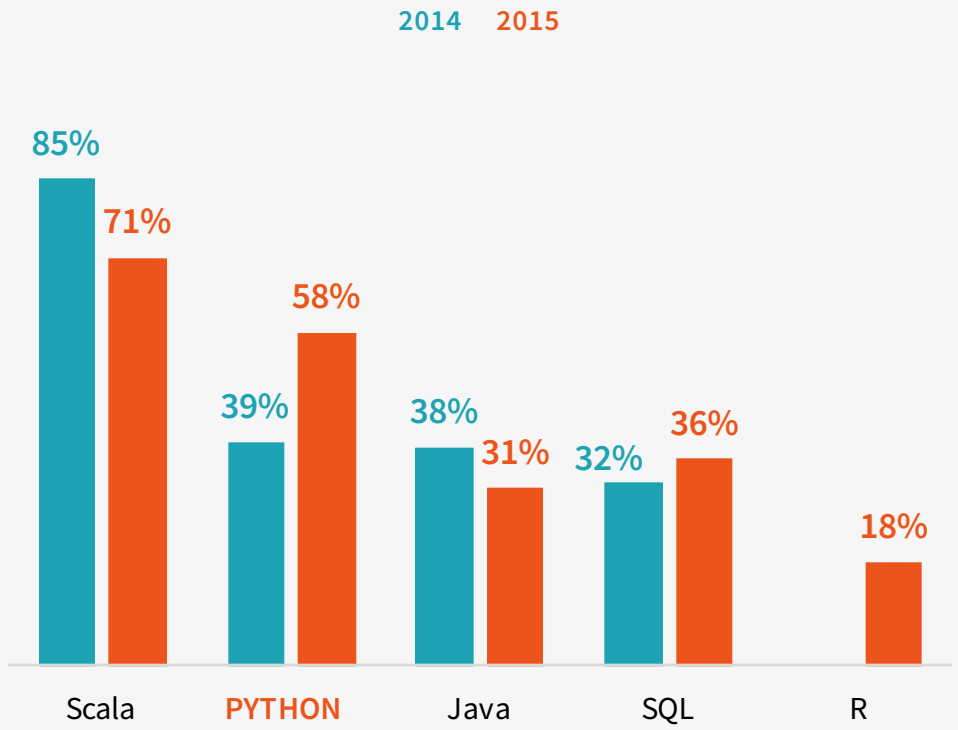
2.4% identify themselves as Other

"These are the two popular languages that people want to do when they do data science. Whether you're a Python person or an R person, Apache Spark is making it simpler and lowering the bar for people to join and talk to their big data."

Ali Ghodsi, Co-Founder and VP of Engineering and Product, Databricks

WHICH LANGUAGES DO YOU USE IN SPARK

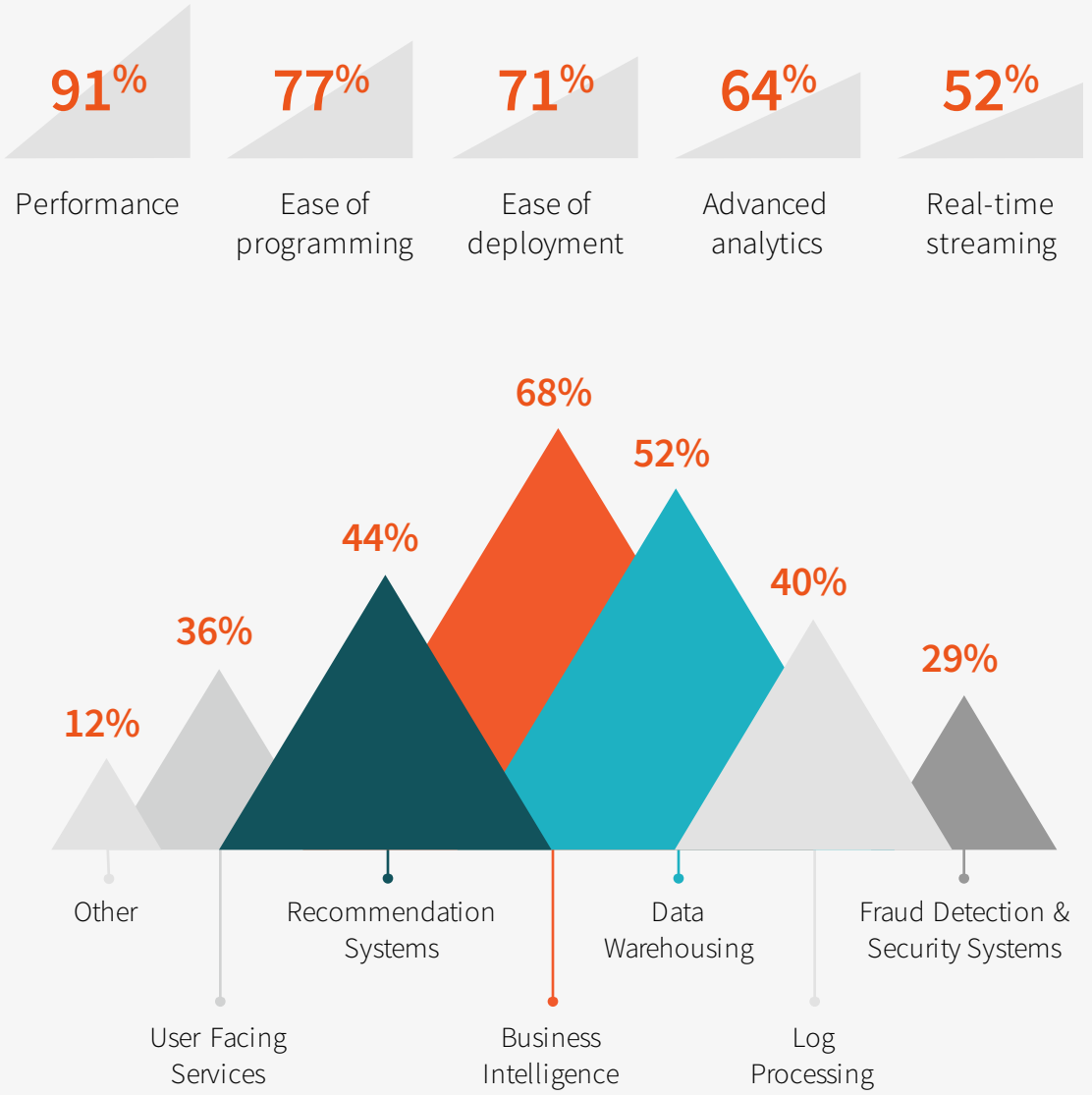
Survey respondents can choose multiple languages.



+49% increase in Python users
(went from 39% to 58% of users)

IN CLOSING...

Spark users are expanding into the areas of advanced analytics and real-time streaming while building foundations on data warehousing and BI.



Conclusion

07

2015 has so far been an exciting year of growth for Spark, and I look forward to the exciting future ahead for the platform.

As Spark expands to new audiences, the ease of use, reliability, and performance that users have come to count on will continue to be a key roadmap priority. With the most recent release of Spark 1.5 and the ongoing Project Tungsten effort, Spark is getting faster and easier to use for anyone who wants to explore data–not just Big Data experts. As advanced analytics and machine learning become higher priorities for Spark's users, we'll continue to focus on the higher level analytics libraries. The transition towards Spark's DataFrames API's is also enabling improvements to the platform that will make it easier for companies to solve data science problems with Spark.

Thanks to the insights revealed in the results of The Spark Survey, we have a better picture of who's using Spark, how they're using it, and what they're using it to build–insights that will guide major updates to the Spark platform as we move into Spark's next phase of growth. Thank you to everyone who participated in The Spark Survey for your help in shaping Spark's future!

Thanks!

PATRICK WENDELL,
Co-founder and engineer at Databricks,
founding Committer and PMC member of Apache Spark
@pwendell

ABOUT

Databricks' vision is to empower anyone to easily build and deploy advanced analytics solutions. The company was founded by the team who created Apache® Spark™, a powerful open source data processing engine built for sophisticated analytics, ease of use, and speed. Databricks is the largest contributor to the open source Apache Spark project providing 10x more code than any other company. The company has also trained over 20,000 users on Apache Spark, and has the largest number of customers deploying Spark to date. Databricks provides a just-in-time data platform, to simplify data integration, real-time experimentation, and robust deployment of production applications. Databricks is venture-backed by Andreessen Horowitz and NEA. For more information, contact info@databricks.com.