databricks®

APACHE® SPARK™ SURVEY 2016 REPORT

APACHE Spark™

Table of Contents

Introduction 3

Foreword: Matei Zaharia 4

REPORT HIGHLIGHTS 5

APACHE SPARK'S GROWTH CONTINUES 13

The Apache Spark Community is Growing 14

Spark's Fastest Growing Areas from 2015 to 2016 17

Spark Users are Growing 18

Spark Users Employ Multiple Languages 19

Spark Components Used in Production 20

Spark is Used Widely in Organizations 21

Users Solve Complex Problems 22

Users Employ Multiple Components 23

What Users Consider Important 24

Top Three Storage Technologies 25

Section Summary 26

APACHE SPARK IN THE CLOUD IS GROWING 27

Trend: Increase in Public Cloud Deployments 28

Trend: Percentage Decrease in On-Premises Deployments 29

Section Summary 30

APACHE SPARK STREAMING AND MACHINE LEARNING SURGE IN USAGE 31

Apache Spark Streaming is Growing 32

Apache Spark Streaming Engine is the Preferred Choice 34

Section Summary 35

Afterword: Reynold Xin 36

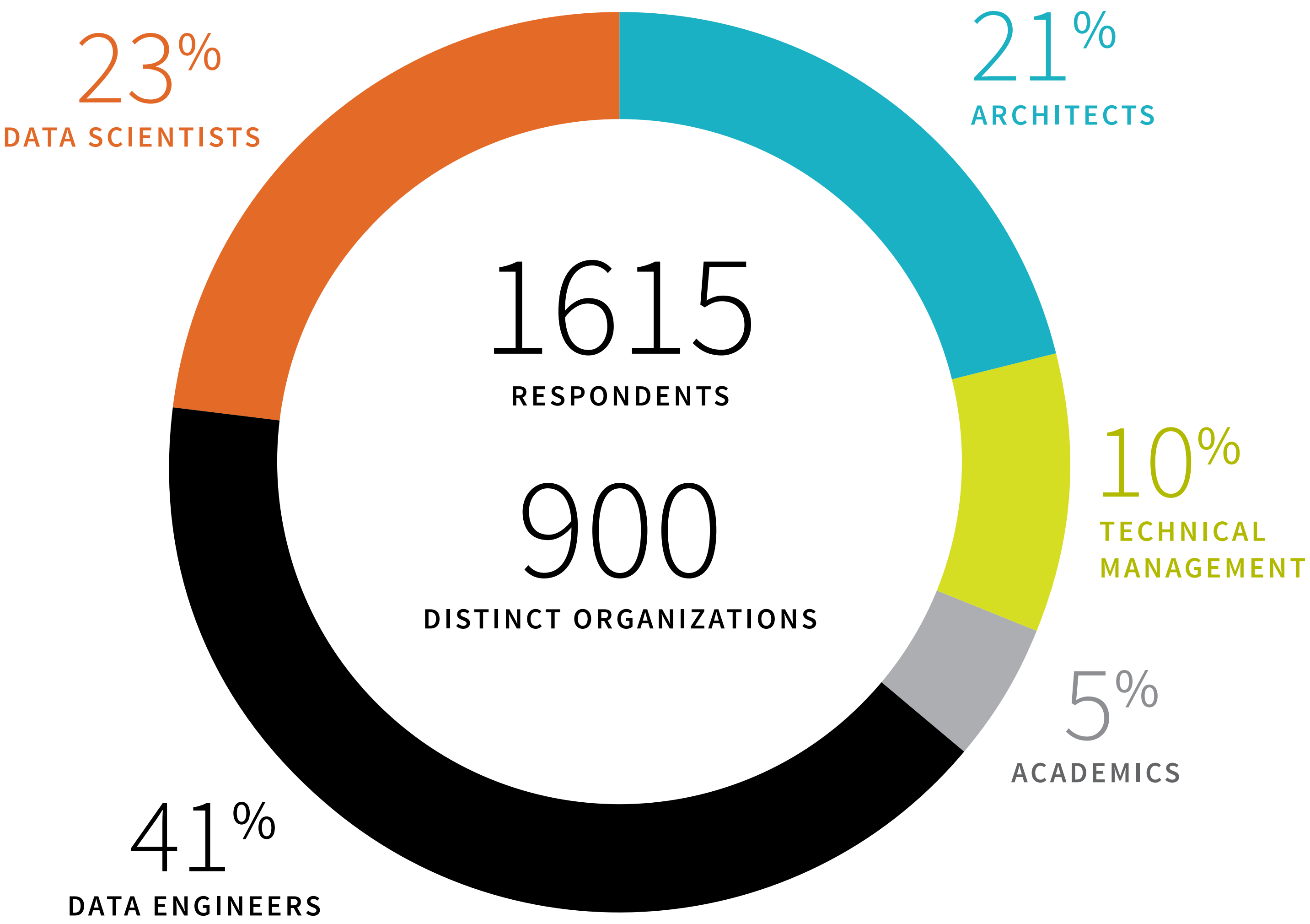
About Databricks 37

Introduction

In July 2016, Databricks conducted an Apache® Spark™ Survey to identify insights into how organizations are using Spark as well as highlight growth trends since the last Spark Survey 2015. In this report, the results reflect answers from over 900 distinct organizations and 1615 respondents, who were predominantly Apache Spark users.

As in 2015, which was a tremendous year in growth for Apache Spark, this year, too, its growth remains unabated—not only in areas like the public cloud, but also with the increased use of Spark Streaming and the use of Machine Learning. 2016 also shows Spark's robust adoption across a variety of organizations and users from many functional roles to build complex solutions, using multiple Spark components. Of the roles represented in the survey, 41% identified themselves as data engineers, while 23% as data scientists and 21% as architects; the rest of the 10% came from technical management and 5% from academia.

SPARK SURVEY 2016



Foreword: Matei Zaharia

I'm delighted to share the results of this year's Databricks Apache Spark Survey. As I noted in the previous Spark Survey 2015, we witnessed a rapid adoption of Spark and the precipitous growth of the Spark community. And this year's Spark's growth trajectory and trends continue. In particular, I'm excited to see more Spark deployments in the cloud and more interest in people building real-time applications using Spark Streaming with multiple components, such as Machine Learning. Given that Apache Spark 2.0 lays the foundational steps for Structured Streaming, by providing simplified and unified APIs to write end-to-end streaming applications called continuous applications, I anticipate this interest will surge further in the coming months—with subsequent releases of Spark.

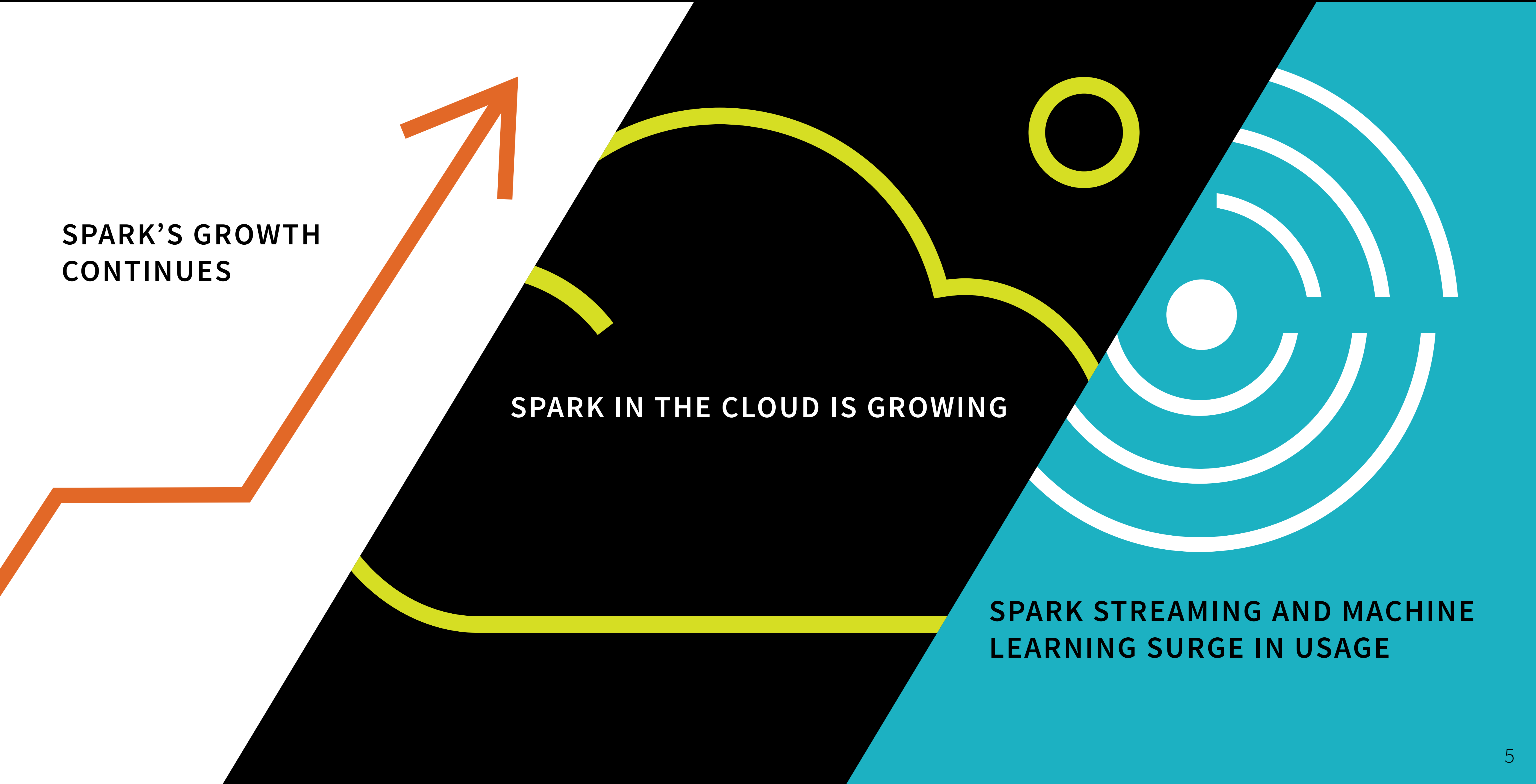
Since its inception, Spark's core mission has been to make Big Data simple and accessible for everyone—for organizations of all sizes and across all industries. And we have not deviated from that mission. In Apache Spark 2.0, we strived to make Spark easier, faster and smarter. And we remain committed to our vision of simplicity. Seventy-six percent of respondents in this survey indicate ease-of-programing as one of the most important features of Spark.

Spark's growth continues across various industries building complex data solutions by people in various functional roles. It has moved well beyond the early-adopter phase at tech companies and is now mainstream in large data-driven enterprises.

> “*Since its inception, Spark's core mission has been to make Big Data simple and accessible for everyone—for organizations of all sizes and across all industries. And we have not deviated from that mission...*”

MATEI ZAHARIA
Chief Technologist at Databricks,
VP of Apache Spark at the Apache Software Foundation
@matei_zaharia

TOP THREE APACHE SPARK TAKEAWAYS



SPARK'S GROWTH CONTINUES

SPARK IN THE CLOUD IS GROWING

SPARK STREAMING AND MACHINE LEARNING SURGE IN USAGE

REPORT HIGHLIGHTS

This year the growth trend continues in the community. Increased growth of Apache Spark Meetup members, a jump in Spark Summit attendees, more code contributors, and a surge in companies represented at the Spark Summit (from several vertical industries) suggest a growing and thriving Spark community.

SPARK MEETUP MEMBERS	CODE CONTRIBUTORS	SPARK SUMMIT ATTENDEES	NUMBER OF COMPANIES AT SUMMITS
+240%	+67%	+30%	+57%
2015 66,000	2015 600	2015 3912	2015 1144
2016 225,000	2016 1000	2016 5100	2016 1800

NOTABLE SPARK USERS WHO PRESENTED AT SPARK SUMMIT 2016

ORACLE Bloomberg YAHOO! Capital One amazon Baidu 百度 airbnb

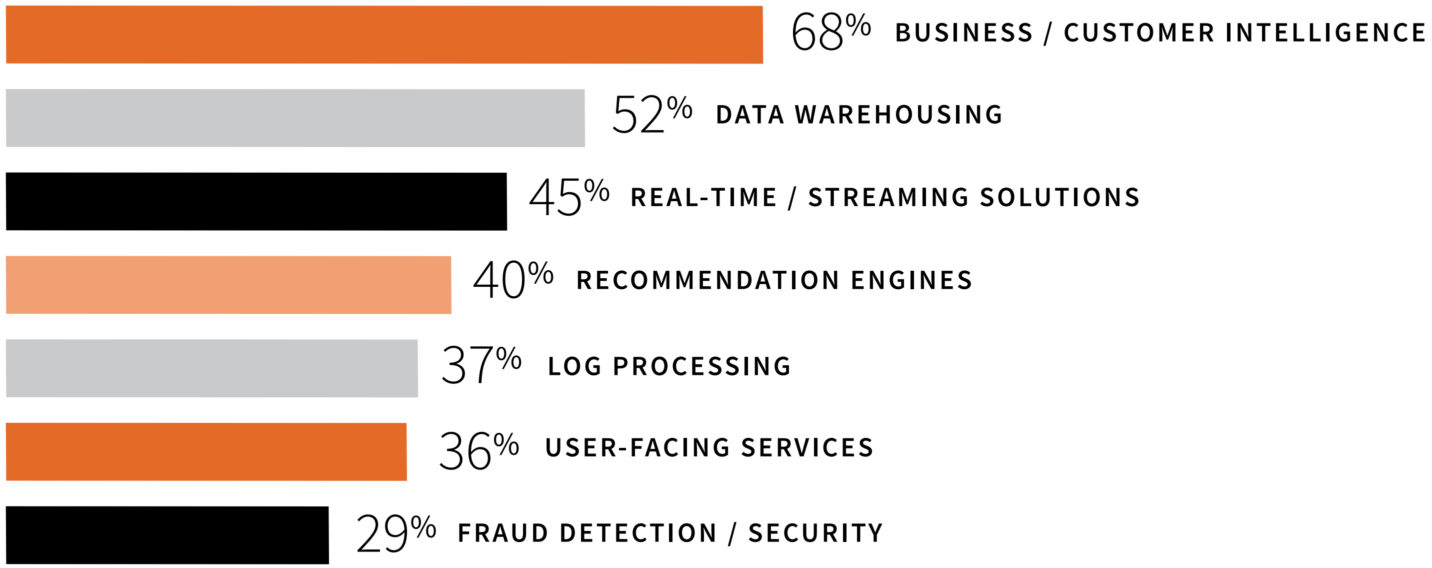
ERICSSON IBM ING intel Google Microsoft NETFLIX

nielsen RIOT GAMES salesforce UBER verizon HUAWEI databricks

Asked what Apache Spark components developers use to build complex solutions for their use cases, **74% of respondents said they use two or more components to build different types of products.**

TYPES OF PRODUCTS BUILT

% of respondents who use Spark to create each product (more than one product could be selected)



NUMBER OF COMPONENTS USED

74%

of respondents

USE TWO OR MORE COMPONENTS

64%

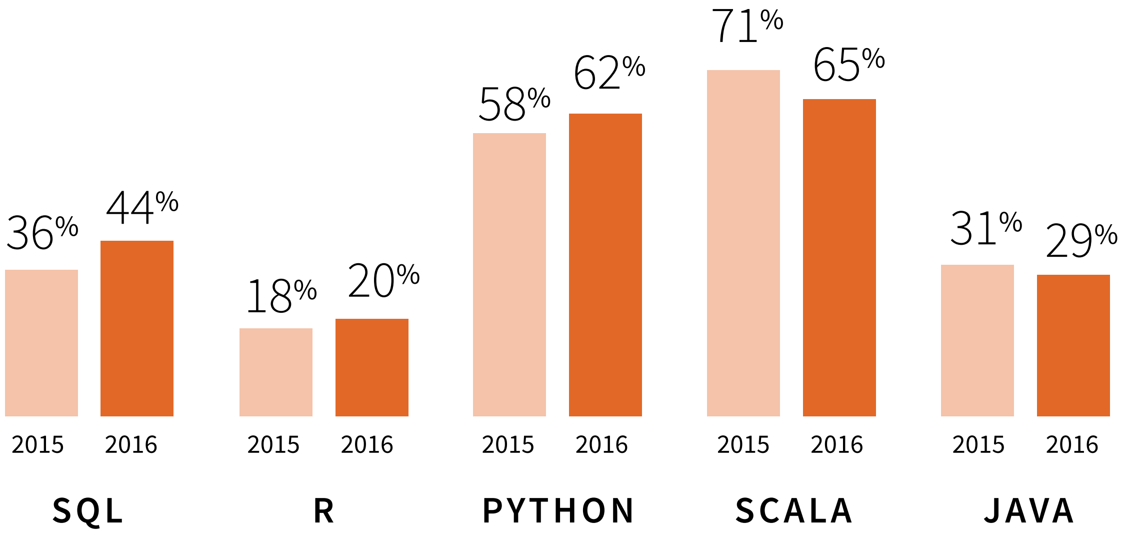
of respondents

USE THREE OR MORE COMPONENTS

In addition to using multiple Apache Spark components, many respondents indicated that they use **multiple programing languages in Spark**. They also are using **multiple components in production**, including **increased use of Spark Streaming and MLlib**.

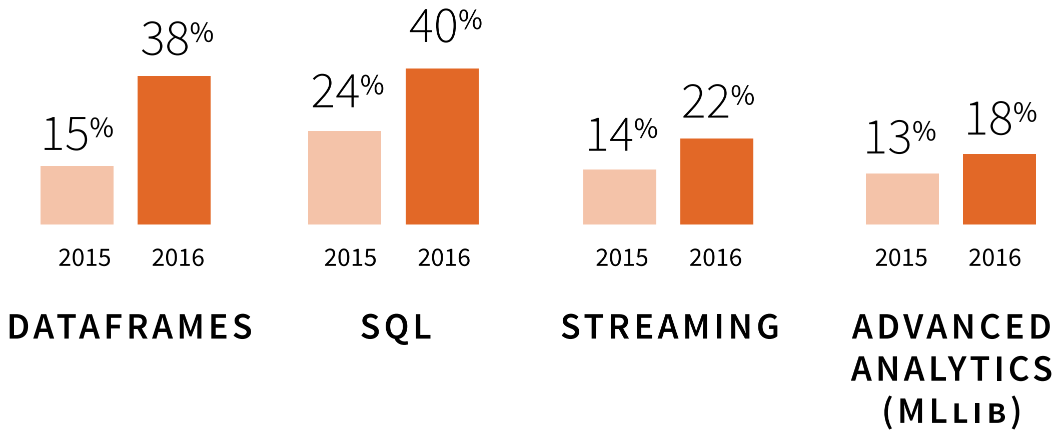
LANGUAGES USED IN SPARK YEAR-OVER-YEAR

% of respondents who use each language (more than one language could be selected)



SPARK COMPONENTS USED IN PRODUCTION YEAR-OVER-YEAR

% of respondents who use each component in production (more than one component could be selected)



APACHE SPARK'S FASTEST GROWING AREAS IN 2016

DATAFRAME* USERS

+153%

2015 15% OF RESPONDENTS

2016 38% OF RESPONDENTS

SPARK SQL* USERS

+67%

2015 24% OF RESPONDENTS

2016 40% OF RESPONDENTS

STREAMING* USERS

+57%

2015 14% OF RESPONDENTS

2016 22% OF RESPONDENTS

ADVANCED ANALYTICS* USERS (MLLIB)

+38%

2015 13% OF RESPONDENTS

2016 18% OF RESPONDENTS

**component used in production*

51% of users in the 2015 Spark Survey said they deployed Apache Spark in the public cloud, compared with 61% of users in 2016, showing a **growth of 20%**.

APACHE SPARK DEPLOYMENT IN PUBLIC CLOUDS INCREASED BY 10% SINCE 2015.

2015

51%

of respondents deployed in a public cloud

→

2016

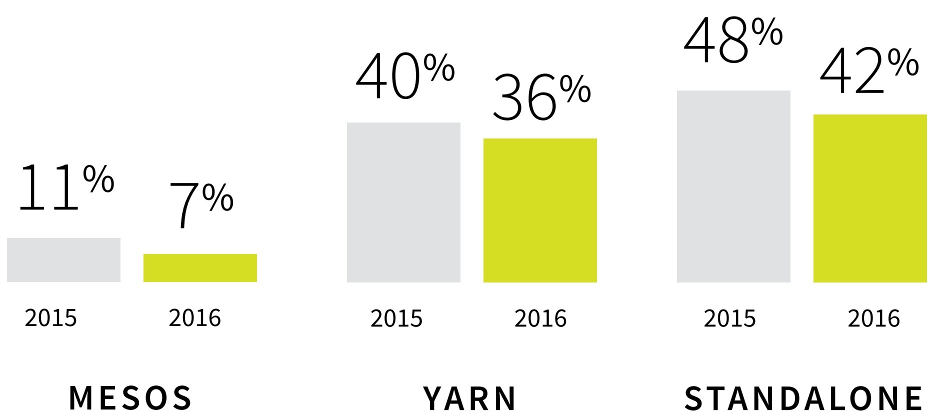
61%

of respondents deploy in a public cloud

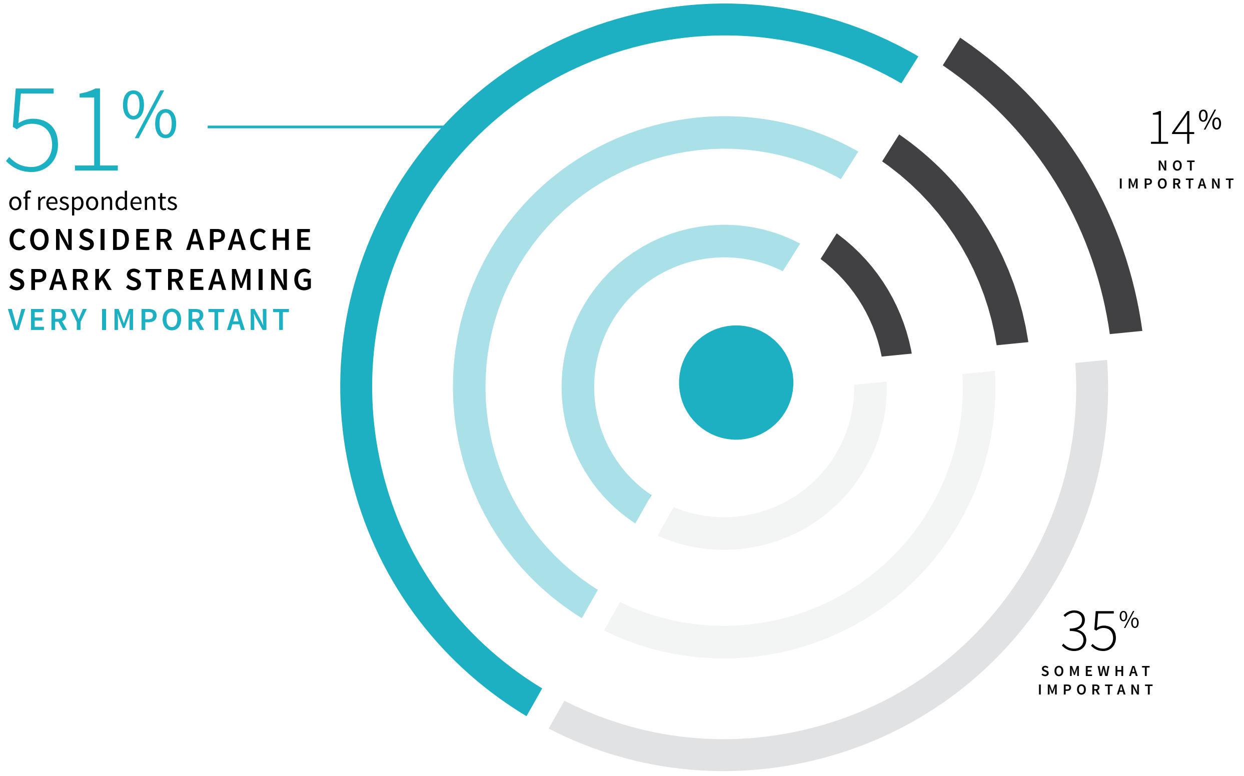
While Apache Spark deployments in the public cloud increased in 2016, **the percentage of Spark deployments on-premises decreased**. For example, 48% of users in 2015 Spark survey and 42% in 2016 survey said they used Standalone cluster managers for their on-premises Spark deployments, showing a 13% percentage decrease. Similarly, YARN and Mesos show 10% and 36% percentage decreases respectively in deployments.

ON-PREMISES DEPLOYMENTS YEAR-OVER-YEAR

% of respondents who use each (more than one deployment could be selected)



Investments in fast data analytics has surged, according to Datanami. Since **companies are shifting investments from batch to real-time applications**, respondents in this survey show an affinity toward building real-time applications using the Spark Streaming framework.



Among all the streaming engines, 33% of respondents said they were heavy users of Spark Streaming.

33% of respondents USE APACHE SPARK STREAMING A LOT

Respondents indicated that Spark Streaming is very important for building real-time streaming, recommendation engines, and fraud detection applications.

Q: WHICH KINDS OF PRODUCTS DOES YOUR ORGANIZATION DEVELOP? *Select all that apply.*

29%

of respondents develop

FRAUD DETECTION / SECURITY PRODUCTS

45%

of respondents develop

REAL-TIME STREAMING PRODUCTS

40%

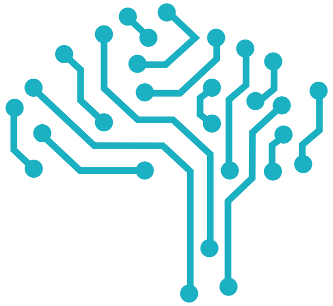
of respondents develop

RECOMMENDATION ENGINE PRODUCTS

Machine Learning has seen an increase in production usage.

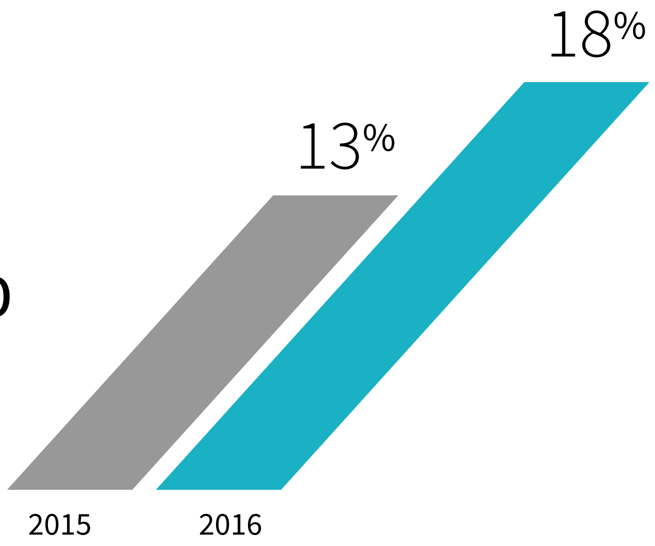
MLLIB USE IN PRODUCTION

% of respondents who use the component in production



+38%

ADVANCED ANALYTICS PRODUCTION CASES



APACHE SPARK'S GROWTH CONTINUES

The Apache Spark Community is Growing

The section identifies key growth areas in all aspects of Spark that are propelling this uptake. Both 2015 and 2016 have seen a tremendous growth in the Spark community and Spark usage in many vertical industries.

Spark today remains the most active open source project in Big Data. Today, there are over 1000 Spark contributors, compared to 600 in 2015 from 250+ organizations. With such large numbers of contributors and organizations investing in Spark's future development, it has engaged a community of developers globally. The Apache Spark Meetup groups' membership continues to flourish, both nationally and internationally.

+67%

CODE CONTRIBUTORS

2015	2016
600	1000

+240%

SPARK MEETUP MEMBERS

2015	2016
66,000	225,000

APACHE SPARK’S GROWTH CONTINUES

Every year, more users attend Spark Summit, the largest dedicated conference to the Apache Spark project. In 2016 there has been an increased number of attendees from a broad range of organizations attending this event, with attendees ranging from developers to data scientists and engineers; to business users and analysts; and executive level decision makers. A number of notable users presented how they use Spark at the Spark Summit San Francisco 2016.

+30%

SPARK SUMMIT ATTENDEES

2015	2016
3912	5100

+57%

COMPANIES REPRESENTED AT SUMMITS

2015	2016
1144	1800

NOTABLE SPARK USERS WHO PRESENTED AT SPARK SUMMIT 2016

ORACLE Bloomberg YAHOO! Capital One amazon

Baidu 百度 airbnb ERICSSON IBM ING

intel Google Microsoft NETFLIX nielsen RIOT GAMES

salesforce UBER verizon HUAWEI databricks

APACHE SPARK’S GROWTH CONTINUES

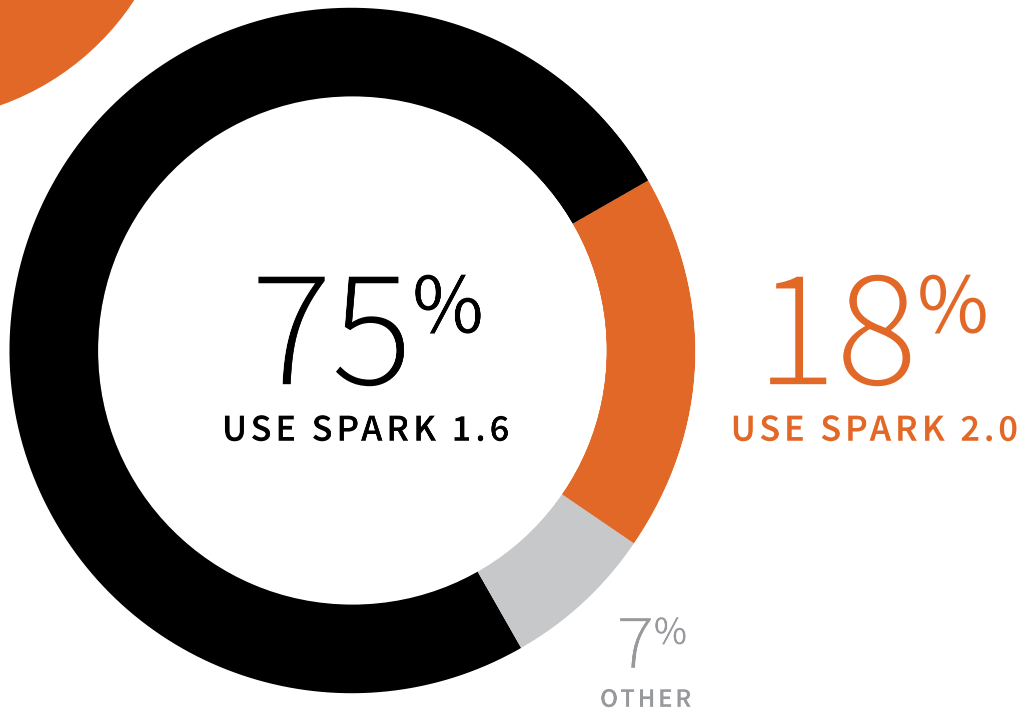
In just two years, the Spark community has released six Spark releases. When asked which version of Apache Spark they are using, 75% responded that they are using Spark 1.6, while 18% are using Spark 2.0 (respondents could choose multiple releases, such as 1.3, or 1.4 or 1.5).

4 RELEASES IN 2015

1.2, 1.3, 1.4, 1.5

2 MAJOR RELEASES IN 2016

1.6, 2.0 *as of September 2016*



Spark's Fastest Growing Areas from 2015 to 2016

Spark Streaming, in particular, has taken a notable increase in its usage, so has SQL, MLlib, and Windows users from 2015.

DATAFRAME USERS IN PRODUCTION

+153%

2015 15% OF RESPONDENTS

2016 38% OF RESPONDENTS

SPARK SQL USERS IN PRODUCTION

+67%

2015 24% OF RESPONDENTS

2016 40% OF RESPONDENTS

STREAMING USERS IN PRODUCTION

+57%

2015 14% OF RESPONDENTS

2016 22% OF RESPONDENTS

WINDOWS USERS IN DEVELOPMENT

+39%

2015 23% OF RESPONDENTS

2016 32% OF RESPONDENTS

ADVANCED ANALYTICS USERS (MLLIB) IN PRODUCTION

+38%

2015 13% OF RESPONDENTS

2016 18% OF RESPONDENTS

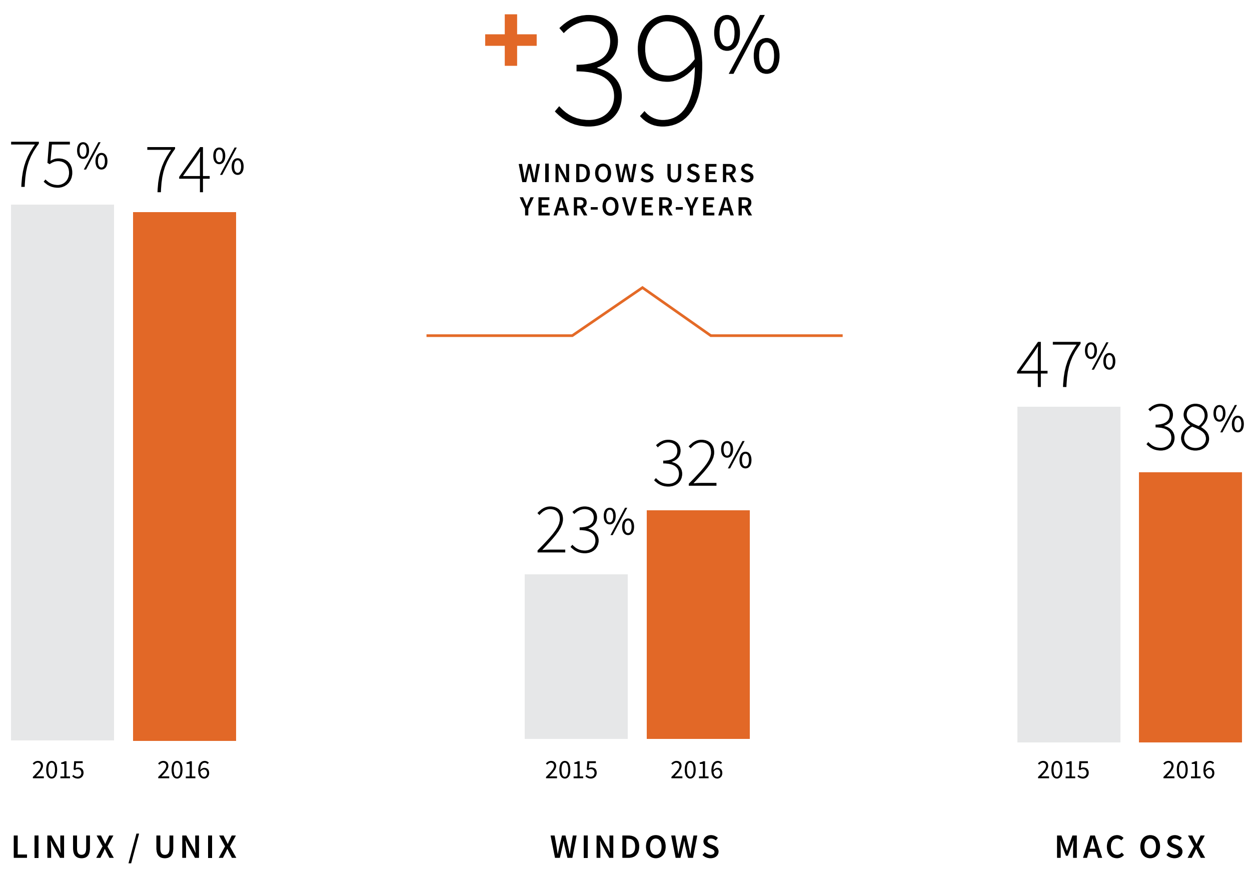
Spark Users are Growing

Spark is attractive not only to highly-skilled and technically advanced users. It crosses barriers, and other users such as **business analysts increasingly use Spark and develop Spark-based applications in environments other than Linux**.

From last year, the percentage of Windows users employing Spark has increased.

DEVELOPMENT ENVIRONMENTS

% of respondents who use each development environment (more than one environment could be selected)



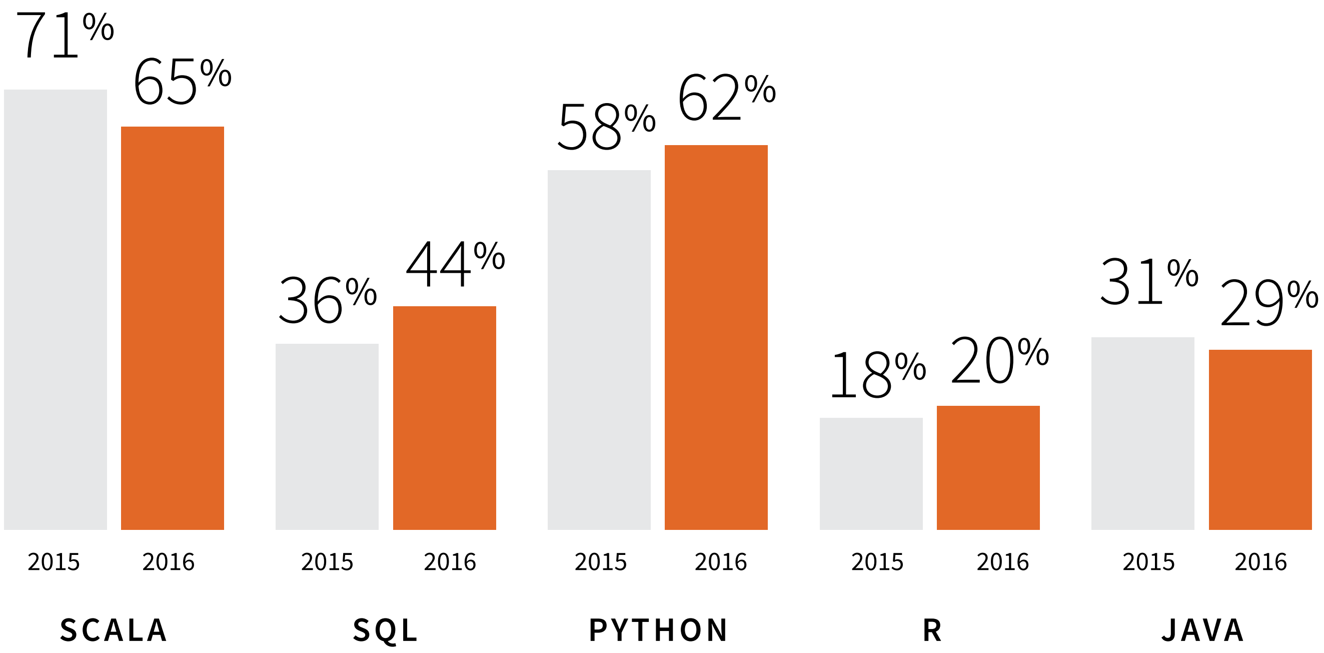
Spark Users Employ Multiple Languages

Spark is becoming the key data processing and computing platform used by a broad range of users. These users span many vertical industries and use a variety of programming languages. One reason for this broad adoption is because **Spark is easy to use and supports familiar programming APIs across these languages**.

Usage of Spark in Python, SQL, and R increased, while Scala and Java usage decreased. This indicates that **more data analysts are drawn to Spark from areas other than pure data engineering**, suggesting that Spark usage is expanding to new and diverse users.

Q: WHICH LANGUAGES DO YOU USE SPARK IN?

% of respondents who use each language (more than one language could be selected)

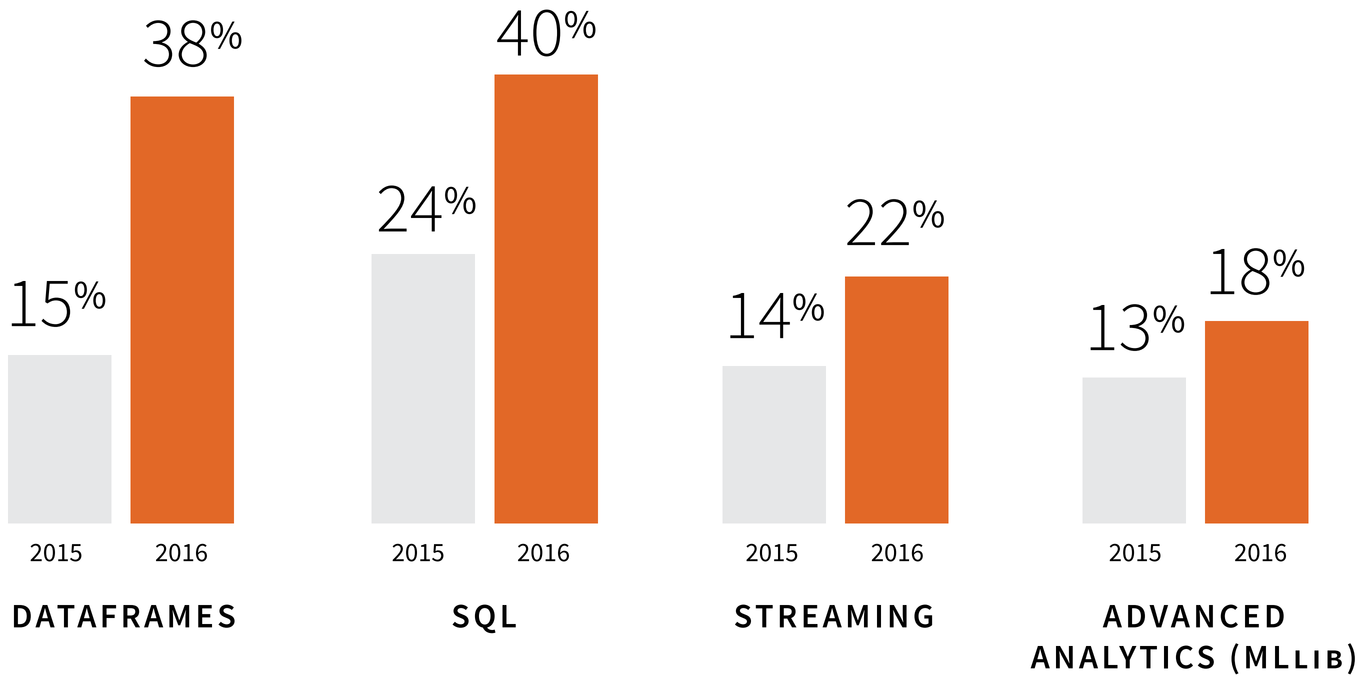


Spark Components Used in Production

Since last year, the use of Spark components in production has increased, especially in Spark Streaming and advanced analytics with Apache Spark MLlib (machine learning). This corroborates with the observation in this report about **increased interest among Spark users to build real-time streaming applications with Spark Streaming, using multiple components, including MLlib**.

Q: WHICH COMPONENTS OF THE APACHE SPARK STACK ARE YOU USING?

% of respondents who use each component ***in production*** *(more than one component could be selected)*



+153% DATAFRAMES USERS

+67% SQL USERS

+57% STREAMING USERS

+38% ADVANCED ANALYTICS USERS

Spark is Used Widely in Organizations

Spark's adoption continues to grow across varied industries because of its unified engine, and because of its proven performance and versatility that enables it to process diverse workloads.

The banking sector saw the highest percentage change in the usage of Spark since 2015, as did the Health, Medical, Biotech and Pharmacy verticals.

+63%
BANKING USERS

2015	2016
6.48%	10.58%

+39%
HEALTH / MEDICAL / PHARMACY / BIOTECH USERS

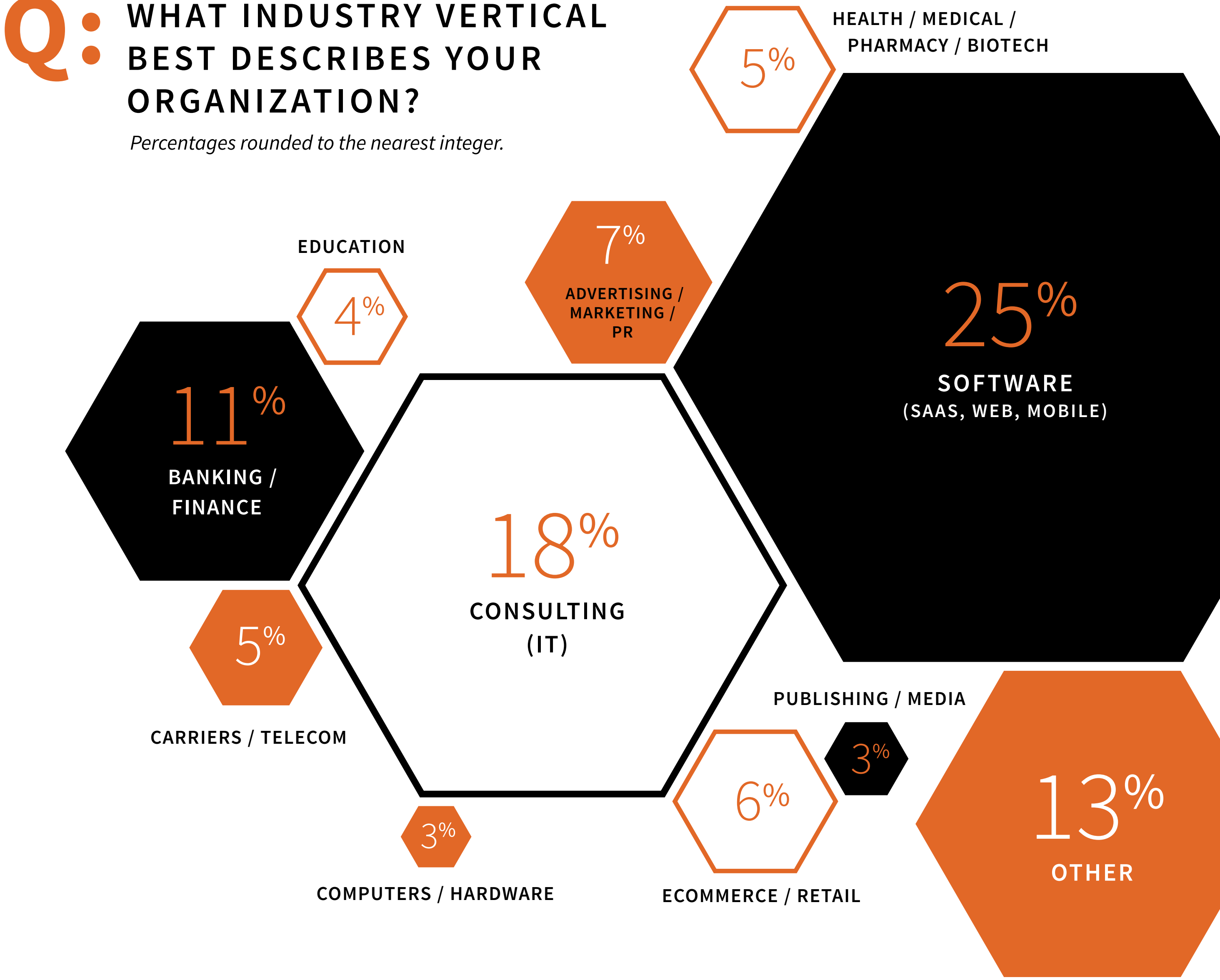
2015	2016
3.89%	5.42%

+29%
CONSULTING (IT) USERS

2015	2016
13.98%	18.09%

Q: WHAT INDUSTRY VERTICAL BEST DESCRIBES YOUR ORGANIZATION?

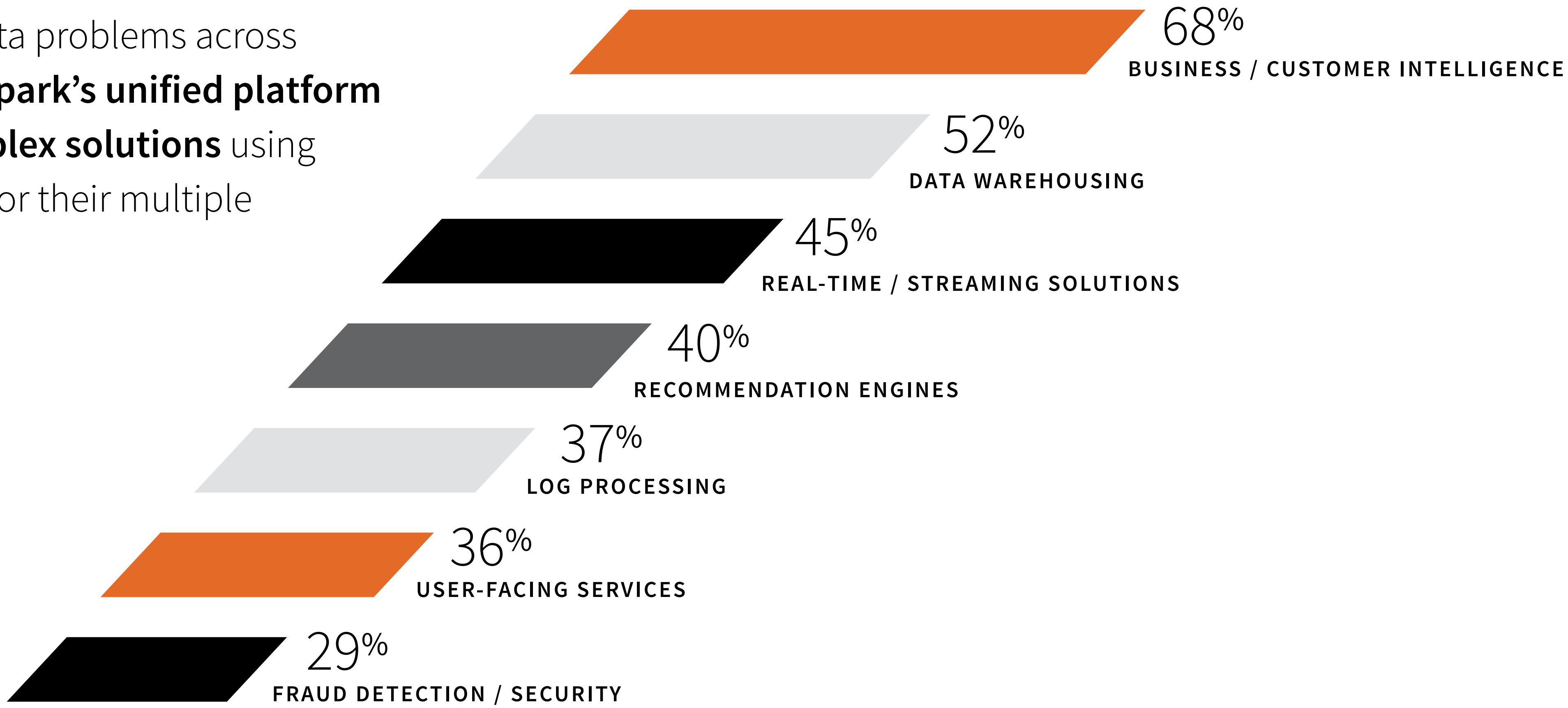
Percentages rounded to the nearest integer.



Users Solve Complex Problems

Users are solving complex data problems across varied industry verticals, as **Spark's unified platform enables users to build complex solutions** using multiple Spark components for their multiple data workloads.

Q: WHICH KINDS OF PRODUCTS DOES YOUR ORGANIZATION DEVELOP? *Select all that apply.*



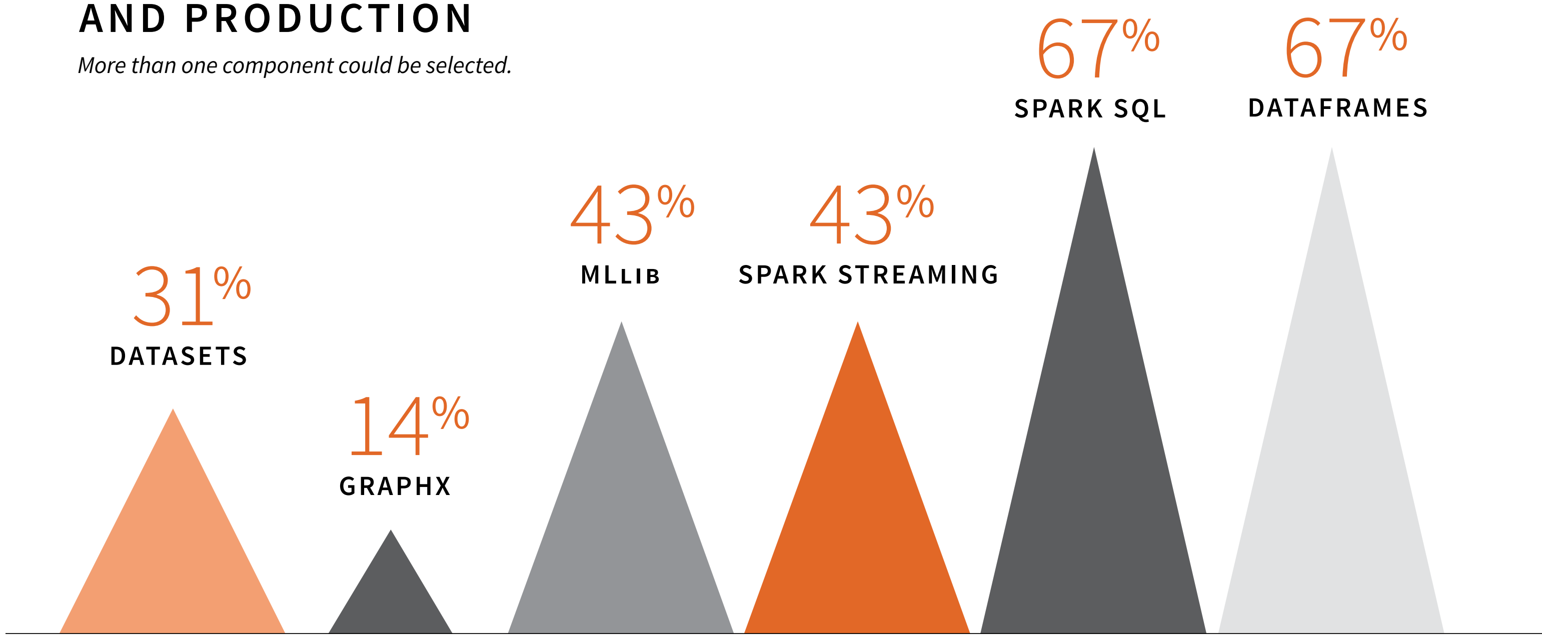
Users Employ Multiple Components

Because of Spark's unified engine and its ability to process multiple workloads within the same cluster, many Spark users within organizations use multiple components of Spark for their use cases and their respective workloads.

Not only are Spark components used separately; **two or more components are often used in prototyping and production**. This unification blurs the barriers between data scientists, data engineers, and data analysts—all using the same unified compute engine.

COMPONENTS USED IN PROTOTYPING AND PRODUCTION

More than one component could be selected.



74%

of Spark users

USE TWO OR MORE COMPONENTS

64%

of Spark users

USE THREE OR MORE COMPONENTS

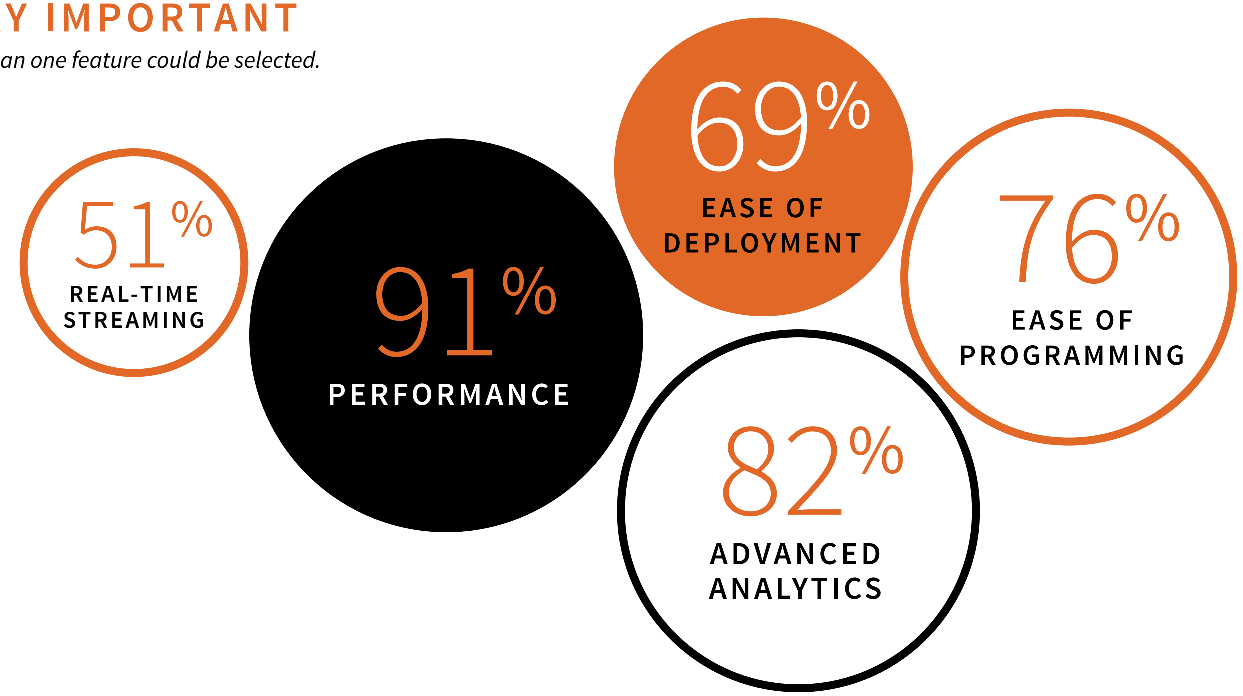
What Users Consider Important

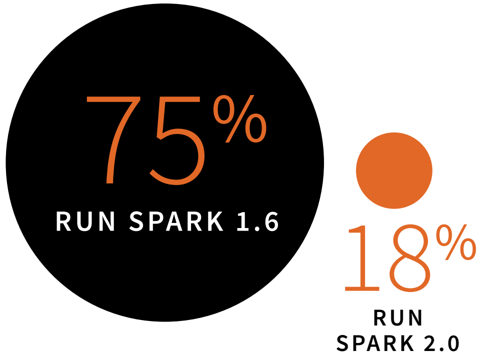
Users are drawn to Spark for a number of reasons: it's easier to get started quickly because of simple and consistent APIs; it's faster because of improvements in Apache Spark 2.0; and it's smarter because of simplified Structured Streaming APIs, allowing users to build end-to-end continuous applications.

According to our 2015 Spark Survey, 91% of users consider performance as the most important aspect of Apache Spark, along with ease of programming, real-time streaming and advanced analytics. In this year's survey, Spark users reflect these as equally important.

% OF RESPONDENTS WHO CONSIDERED THE FEATURE VERY IMPORTANT

More than one feature could be selected.





At the time of this survey, Apache Spark 2.0 had just been officially released, and users displayed a keen interest in using it. Even though most users run Spark 1.6, the 2016 survey results suggest they had quickly started using Spark 2.0.

Top Three Storage Technologies

A large number of Spark users use technologies for storage other than Apache® Hadoop®, such as Cassandra, MongoDB and NoSQL as well as other open-source and proprietary SQL data stores.

Q: WHICH OF THESE TECHNOLOGIES DO YOU CURRENTLY USE? *Select all that apply.*

82%

of respondents use

OPEN-SOURCE SQL DATABASES

73%

of respondents use

KEY-VALUE STORES (NoSQL)

58%

of respondents use

PROPRIETARY SQL DATABASES

Section Summary

Apache Spark's growth and adoption continues as users, industries, development environments, disciplines, and programming languages embrace its ease of use and programming, its unified compute engine, and its performance to solve complex data problems at scale. Spark allows multiple components to work on multiple workloads and access data from multiple data sources. All of these factors make Spark an attractive choice as a unified compute data platform.

APACHE SPARK IN THE CLOUD IS GROWING

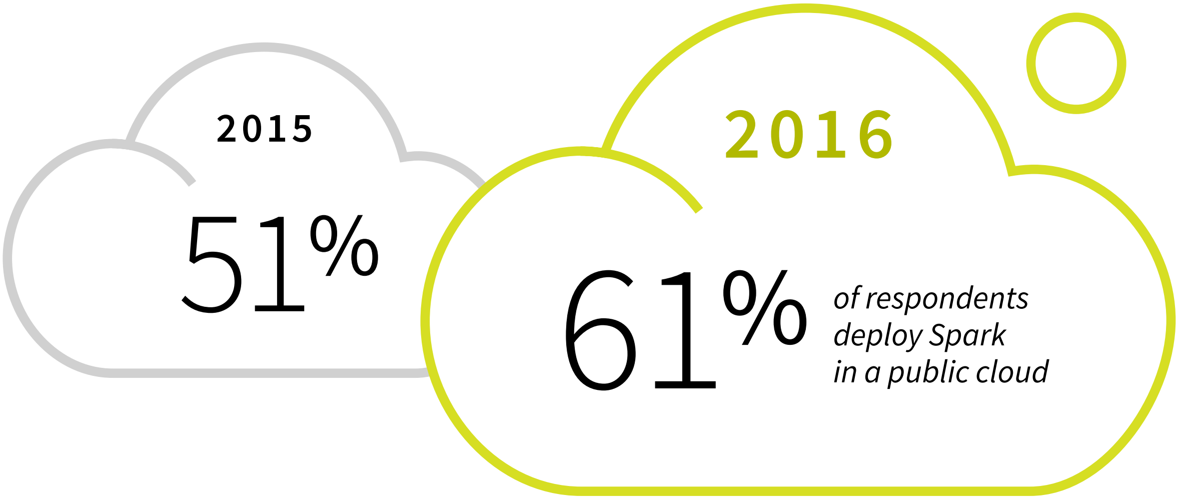
Trend: Increase in Public Cloud Deployments

The rise of cloud computing is rapid, inexorable and causing a huge upheaval in the tech industry, writes The Economist. "Gartner estimates that about $205 billion, or 6% of the world's IT budget of $3.4 trillion, will be spent on cloud computing in 2016—a number it expects to grow to $240 billion next year," according to another article in The Economist.

This survey reflects this trend, as **many respondents are electing to deploy Spark in the public cloud, mitigating both cost and infrastructure headaches.**

Since 2015, we have seen a 20% growth of users deploying Spark in the public cloud. That is, 61% users in the 2016 survey said they deployed Spark in the public cloud compared to 51% in 2015.

SPARK DEPLOYMENT IN PUBLIC CLOUDS HAS INCREASED BY 10% SINCE 2015.

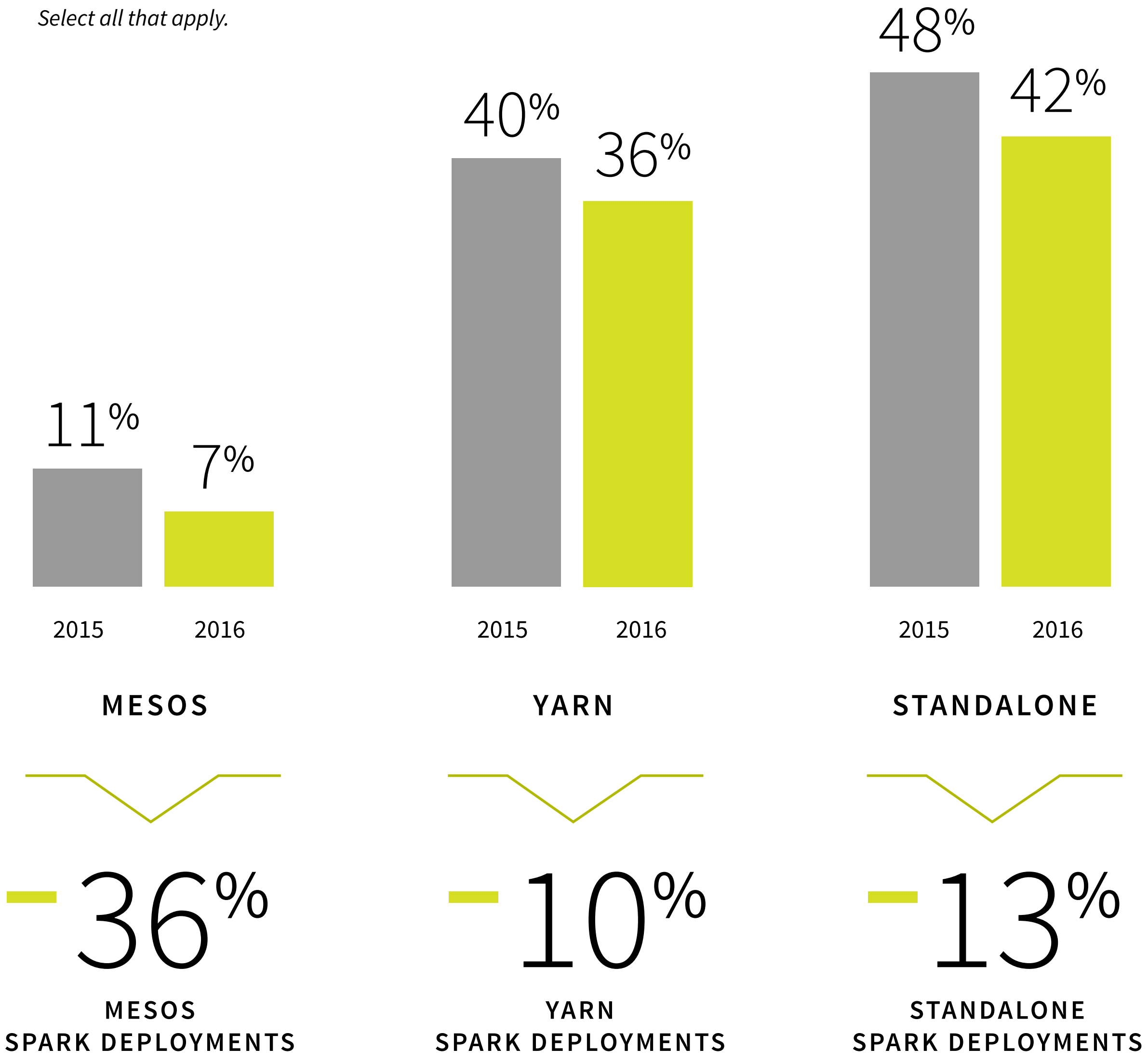


Trend: Percentage Decrease in On-Premises Deployments

Although many Spark users run Spark on-premises alongside Hadoop and other data sources, **some deployment modes in 2016 have seen a percentage decrease**.

Q: WHERE DO YOU RUN SPARK?

Select all that apply.



Section Summary

Not only do cloud deployments have lower deployment costs and fewer management headaches, they have higher and proven performance benefits.

> “*Using Apache Spark on 206 EC2 machines, we sorted 100TB 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
Chief Architect & Co-Founder of Databricks

APACHE SPARK STREAMING AND MACHINE LEARNING SURGE IN USAGE

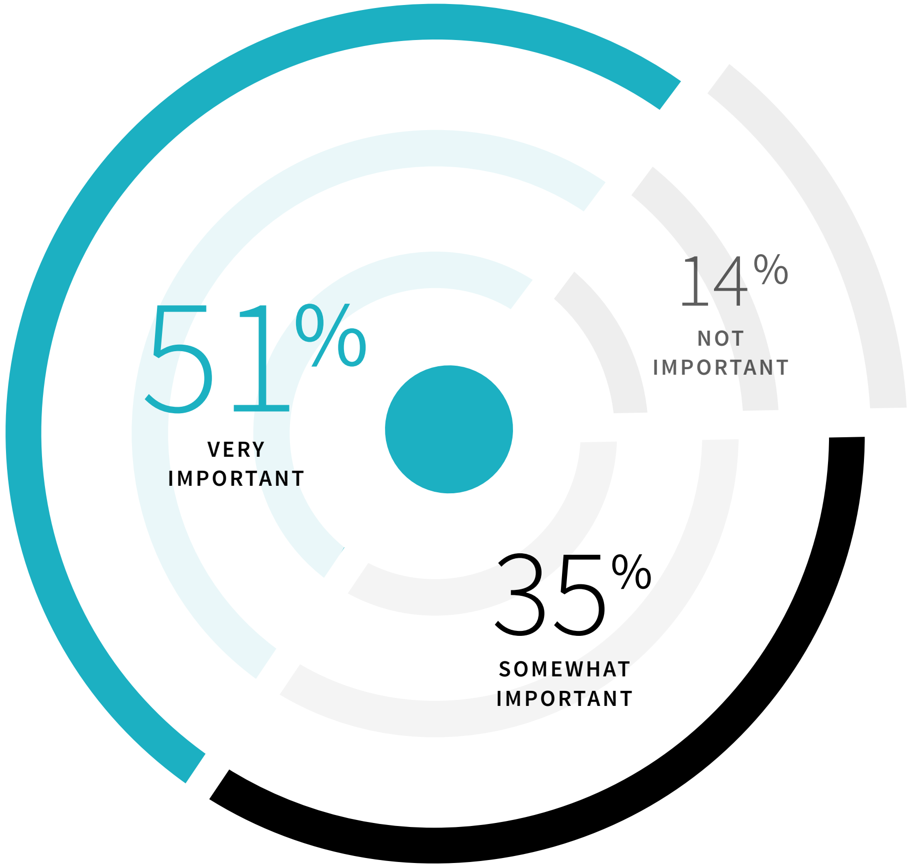
Apache Spark Streaming is Growing

Since its release, Spark Streaming has become one of the most widely used distributed streaming engines. Interest in developing real-time applications and advanced analytics is on the rise.

Over half of the survey respondents indicate that streaming is vital and important for developing valuable real-time streaming, recommendation engines, and fraud-detection and security solutions.

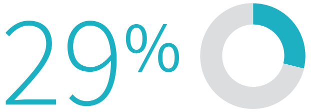
Q:

HOW IMPORTANT IS SPARK STREAMING TO YOUR USE CASE?

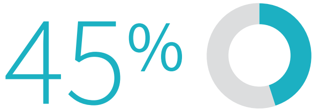


Q:

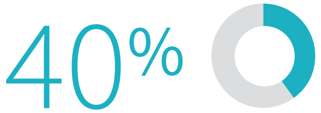
WHICH KINDS OF PRODUCTS DOES YOUR ORGANIZATION DEVELOP? *Select all that apply.*



of respondents develop
FRAUD DETECTION / SECURITY PRODUCTS



of respondents develop
REAL-TIME STREAMING PRODUCTS



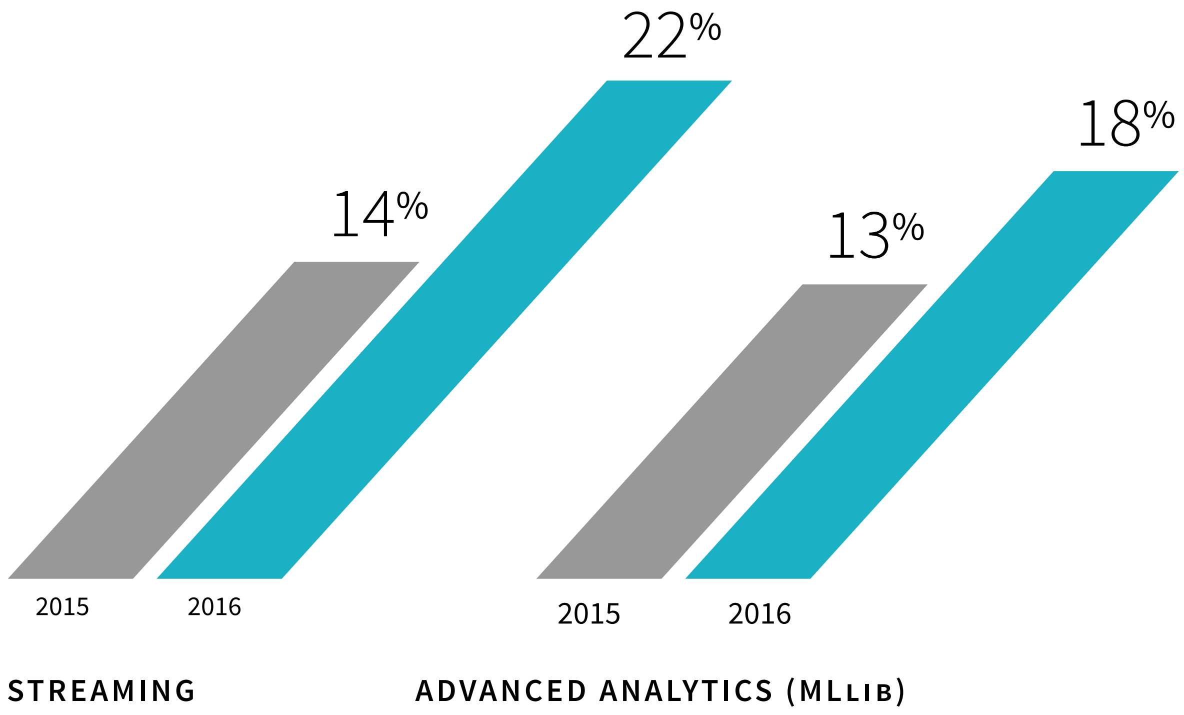
of respondents develop
RECOMMENDATION ENGINE PRODUCTS

APACHE SPARK STREAMING AND MACHINE LEARNING SURGE IN USAGE

Organizations use Spark Streaming along with Spark's other multiple components to develop streaming applications. Both Spark Streaming and MLlib saw a notable increase in production use.

SPARK STREAMING AND MLLIB USE IN PRODUCTION

% of respondents who use the component in production (more than one component could be selected)



+57%

STREAMING PRODUCTION CASES

+38%

ADVANCED ANALYTICS PRODUCTION CASES

APACHE SPARK STREAMING AND MACHINE LEARNING SURGE IN USAGE

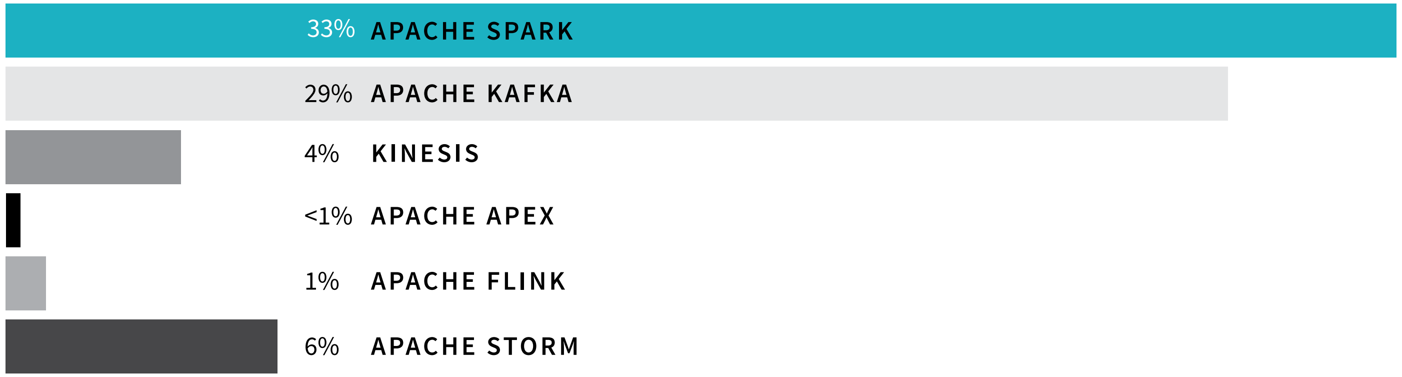
Apache Spark Streaming Engine is the Preferred Choice

Compared to other streaming engines, Spark is the preferred choice at 33%.

When compared to other Spark components, Spark Streaming matches MLlib at 71% in use, from evaluation to production.

In the 2015 Spark survey, 14% of users said they used Spark Streaming in production, compared to 22% of users in 2016. Overall, we saw a 57% growth of users using Spark Streaming in production.

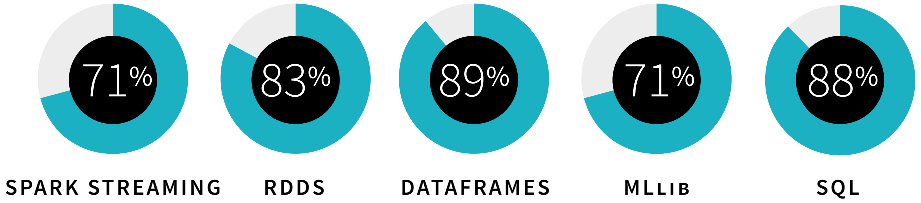
Q: WHICH OF THESE TECHNOLOGIES DO YOU CURRENTLY USE A LOT FOR STREAMING AND/OR COMPLEX EVENT PROCESSING CASES? *Select all that apply.*



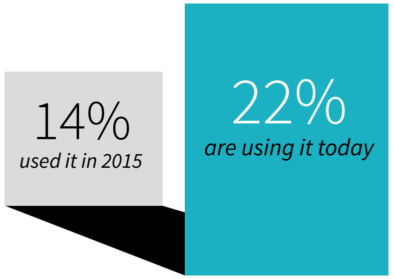
Note: Respondents were predominately Spark users.

APACHE SPARK COMPONENT POPULARITY

% of respondents who use the component anywhere from evaluation to production (more than one component could be selected)



Q: DO YOU CURRENTLY USE SPARK STREAMING IN PRODUCTION?



+57%

SPARK STREAMING PRODUCTION CASES

Section Summary

Spark Streaming is being used for real-time solutions, from evaluation to production, closer in usage to Spark's other commonly used components. As a preferred choice of streaming engine over others, more organizations are building real-time streaming solutions as they consider streaming an important Spark feature.

Afterword: Reynold Xin

2015 and 2016 have been exciting years for the adoption and increased growth of Apache Spark and its community. Two releases—Spark 1.6 and 2.0—have seen major improvements in all aspects of Spark noted by respondents in this survey as important. I continue to look forward, and work with the community, to the exciting future ahead for the Spark platform.

As Spark becomes easier, faster, and smarter, outside the predominantly IT and Consulting Industry, a newer audience is adopting it, as results from the survey suggest. Performance, ease-of-use, streaming, and reliability top the list as most important features. At the time of this survey, we released Apache Spark 2.0. Ongoing performance improvements, with Project Tungsten, started in earlier releases and culminated in Spark 2.0. In addition, Spark 2.0 delivered unified DataFrames and Datasets APIs and simplified Structured Streaming APIs. All these make Spark an attractive engine for performing advanced analytics across industry verticals in solving complex data problems, by users from different functional roles.

Your voice matters. We got an insightful glimpse into the growth and trends from this year's survey: who's using Spark, how they are using it, what's important, what new features they use, and what they are using it for. Just as the feedback from last year's survey did, these insights will drive major updates and help shape the future of the Spark platform.

Thank you to everyone who participated in Databricks' Apache Spark Survey 2016!

REYNOLD XIN
Chief Architect & Co-Founder of Databricks
@rxin

databricks

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.

TRY DATABRICKS FOR FREE
databricks.com/try-databricks

CONTACT US FOR A PERSONALIZED DEMO
databricks.com/contact-databricks