Spark Survey Results

2015

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 over 1,417 respondents representing over 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. The survey also indicates that Spark is increasingly used outside of Hadoop environments – a revelation that promises an exciting future for Spark.

1. Spark Adoption Is Growing Rapidly

Adoption of Spark has spread beyond the technology industry, and Spark is fast becoming the Big Data technology for everyone, not just for Big Data experts.

SPARK IS THE MOST ACTIVE OPEN SOURCE PROJECT IN BIG DATA.

Spark Summit conferences

(based on Spark Summit East and Spark Summit West, not including Spark Summit Europe)

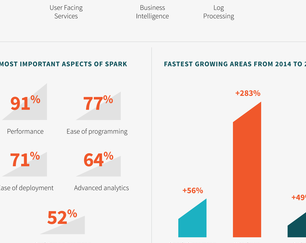
2014	2015*
1,164 attendees	2,986 attendees
453 companies	1,144 companies

Spark contributors

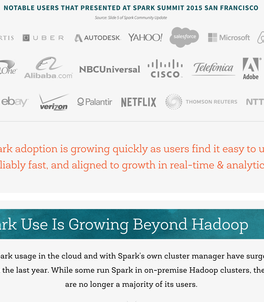
600 Last 12 months

315 Last 12-24 months

TOP 10 INDUSTRIES USING SPARK



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



MOST IMPORTANT ASPECTS OF SPARK

91% Performance

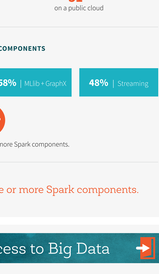
77% Ease of programming

71% Ease of deployment

64% Advanced analytics

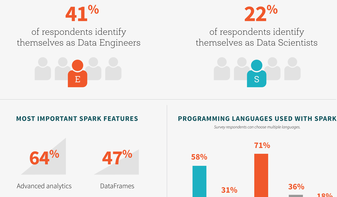
52% Real-time streaming

FASTEST GROWING AREAS FROM 2014 TO 2015



NOTABLE USERS THAT PRESENTED AT SPARK SUMMIT 2015 SAN FRANCISCO

Source: Slide 3 of Spark Community Update



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

2. Spark Use Is Growing Beyond Hadoop

Spark usage in the cloud and with Spark's own cluster manager have surged in the last year. While some run Spark in on-premise Hadoop clusters, they are no longer a majority of its users.

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.

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

3. Spark Is Increasing Access to Big Data

Spark is unlocking the value of Big Data by making it easier for a wide range of people to solve a growing variety of data problems.

TOP ROLES USING SPARK

41% of respondents identify themselves as Data Engineers

22% of respondents identify themselves as Data Scientists

MOST IMPORTANT SPARK FEATURES

64% Advanced analytics

47% DataFrames

52% Real-time streaming

28% SQL Standards

PROGRAMMING LANGUAGES USED WITH SPARK

(Survey respondents can choose more than one language)

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

Feedback from the Spark community is vital in planning major updates to the Spark platform. Thank you to all the respondents of the 2015 Spark Survey for helping shape the future of Spark. Dive deeper into the Spark Survey in the Spark Survey Report 2015.

ABOUT

databricks

Databricks' vision is to dramatically simplify big data processing. It was founded by the team that created and continues to drive Apache Spark, a powerful open source data processing engine built for sophisticated analytics, ease of use, and speed. Databricks offers a cloud-based integrated workspace for big data that lets users go from data ingest, to visual exploration and production jobs, making it easy to turn data into value, without the hassle of managing complex infrastructure, systems and tools. Databricks is venture-backed by Andreessen Horowitz and NEA. For more information, contact info@databricks.com.