2017 Edition

Big Data Analytics Market Study

Wisdom of Crowds Series

Licensed to MicroStrategy
Disclaimer:

This report should be used for informational purposes only. Vendor and product selections should be made based on multiple information sources, face-to-face meetings, customer reference checking, product demonstrations, and proof-of-concept applications.

The information contained in all Wisdom of Crowds® Market Study Reports reflects the opinions expressed in the online responses of individuals who chose to respond to our online questionnaire and does not represent a scientific sampling of any kind. Dresner Advisory Services, LLC shall not be liable for the content of reports, study results, or for any damages incurred or alleged to be incurred by any of the companies included in the reports as a result of its content.

Reproduction and distribution of this publication in any form without prior written permission is forbidden.
Definition

**Big Data Analytics Defined**
We define big data analytics as systems that enable end-user access to and analysis of data contained and managed within the Hadoop ecosystem.
Introduction
This year we celebrate the tenth anniversary of Dresner Advisory Services and our first-ever conference, Real Business Intelligence, which took place July 11-12 on the MIT campus in Cambridge, Massachusetts!

Our thanks to all of you for your continued support and ongoing encouragement!

Since our founding in 2007, we have worked hard to set the “bar” high—challenging ourselves to innovate and lead the market—offering ever greater value with each successive year.

Our first market report in 2010 set the stage for where we are today. Since that time, we expanded our agenda and added new research topics every year. For 2017, we plan to release 15 major reports, including our original BI flagship report—in its eighth year of publication.

In previous years, we added new topics to our agenda, and 2017 is no exception. Earlier this year, we published our inaugural Analytical Data Infrastructure report and added data catalog to the existing lineup during Q2.

For this, our third Big Data Analytics Market Study, we continue to focus upon the combination of analytical solutions within the Hadoop ecosystem, adding some new criteria and exploring changing market dynamics and user perceptions and plans.

We hope you enjoy this report!

Best,

Howard Dresner
Chief Research Officer
Dresner Advisory Services
# Contents

Definition .......................................................................................................................... 3

Big Data Analytics Defined .............................................................................................. 3

Introduction ....................................................................................................................... 4

Benefits of the Study ........................................................................................................ 7

A Consumer Guide ........................................................................................................... 7

A Supplier Tool .................................................................................................................. 7

About Howard Dresner and Dresner Advisory Services .................................................. 8

About Jim Ericson .............................................................................................................. 9

Survey Method and Data Collection ............................................................................... 10

Data Quality ..................................................................................................................... 10

Executive Summary ......................................................................................................... 12

Study Demographics ....................................................................................................... 13

Geography ....................................................................................................................... 13

Functions .......................................................................................................................... 14

Vertical Industries ........................................................................................................... 15

Organization Size ............................................................................................................ 16

Analysis and Trends: Big Data Analytics ....................................................................... 18

Importance of Big Data .................................................................................................... 18

Big Data Adoption ............................................................................................................ 19

Future Adoption of Big Data .......................................................................................... 25

Big Data Use Cases ......................................................................................................... 31

Big Data Infrastructure .................................................................................................... 37

Big Data – Data Access .................................................................................................... 43

Big Data Search ................................................................................................................ 49

Big Data Analytics / Machine-Learning Technologies ..................................................... 55

Big Data Distributions .................................................................................................... 61

Industry and Vendor Analysis .......................................................................................... 68

Big Data Analytics Vendor Ratings ................................................................................ 79

Vendor Ratings Detail ....................................................................................................... 80
Top Vendors for Big Data Infrastructure ................................................................. 80
Top Vendors for Big Data / Data Access ............................................................... 80
Top Vendors for Big Data Search ..................................................................... 80
Top Vendors for Big Data Machine Learning .................................................. 80
Top Vendors for Big Data Distribution Support ............................................. 80
Glossary ........................................................................................................... 81
Other Dresner Advisory Services Research Reports ........................................ 85
Appendix: Big Data Analytics Study Survey Instrument .................................. 86
Benefits of the Study
The DAS Big Data Analytics Market Study provides a wealth of information and analysis, offering value to both consumers and producers of related technology and services.

A Consumer Guide
As an objective source of industry research, consumers use the DAS Big Data Analytics Market Study to understand how their peers are leveraging and investing in big data analytics and related technologies.

Using our unique vendor performance measurement system, users glean key insights into software supplier performance, enabling:

- Comparisons of current vendor performance to industry norms
- Identification and selection of new vendors

A Supplier Tool
Vendor licensees use the DAS Big Data Analytics Market Study in several important ways:

External Awareness

- Build awareness for the big data analytics market and supplier brand, citing DAS Big Data Analytics Market Study trends and vendor performance
- Create lead and demand generation for supplier offerings through association with DAS Big Data Analytics Market Study brand, findings, webinars, etc.

Internal Planning

- Refine internal product plans and align with market priorities and realities as identified in DAS Big Data Analytics Market Study
- Better understand customer priorities, concerns, and issues
- Identify competitive pressures and opportunities
About Howard Dresner and Dresner Advisory Services

The DAS Big Data Analytics Market Study was conceived, designed, and executed by Dresner Advisory Services, LLC, an independent advisory firm, and Howard Dresner, its president, founder and chief research officer.

Howard Dresner is one of the foremost thought leaders in business intelligence and performance management, having coined the term “Business Intelligence” in 1989. He has published two books on the subject, *The Performance Management Revolution – Business Results through Insight and Action* (John Wiley & Sons, Nov. 2007) and *Profiles in Performance – Business Intelligence Journeys and the Roadmap for Change* (John Wiley & Sons, Nov. 2009). He lectures at forums around the world and is often cited by the business and trade press.

Prior to Dresner Advisory Services, Howard served as chief strategy officer at Hyperion Solutions and was a research fellow at Gartner, where he led its business intelligence research practice for 13 years.

Howard has conducted and directed numerous in-depth primary research studies over the past two decades and is an expert in analyzing these markets.

Through the Wisdom of Crowds® Business Intelligence market research reports, we engage with a global community to redefine how research is created and shared. Other research reports include:

- Advanced and Predictive Analytics
- Business Intelligence Competency Center
- Cloud Computing and Business Intelligence
- Collective Insights®
- Embedded Business Intelligence
- End User Data Preparation
- IoT Intelligence®
- Location Intelligence

Howard conducts a weekly Twitter “tweetchat” on Fridays at 1:00 p.m. ET. During these live events the #BIWisdom “tribe” discusses a wide range of business intelligence topics.

You can find more information about Dresner Advisory Services at www.dresneradvisory.com.
About Jim Ericson

Jim Ericson is Vice President and Research Director with Dresner Advisory Services.

Jim has served as a consultant and journalist who studies end-user management practices and industry trending in the data and information management fields.

From 2004 to 2013 he was the editorial director at Information Management magazine (formerly DM Review), where he created architectures for user and industry coverage for hundreds of contributors across the breadth of the data and information management industry.

As lead writer, he interviewed and profiled more than 100 CIOs, CTOs, and program directors in a 2010-2012 program called “25 Top Information Managers.” His related feature articles earned ASBPE national bronze and multiple Mid-Atlantic region gold and silver awards for Technical Article and for Case History feature writing.

A panelist, interviewer, blogger, community liaison, conference co-chair, and speaker in the data-management community, he also sponsored and co-hosted a weekly podcast in continuous production for more than five years.

Jim’s earlier background as senior morning news producer at NBC/Mutual Radio Networks and as managing editor of MSNBC’s first Washington, D.C. online news bureau cemented his understanding of fact-finding, topical reporting, and serving broad audiences.
Survey Method and Data Collection
As with all of our Wisdom of Crowds® Market Studies, we constructed a survey instrument to collect data and used social media and crowd-sourcing techniques to recruit participants.

We include our own research community of nearly 4,000 organizations as well as crowdsourcing and vendors’ customer communities.

Data Quality
We carefully scrutinized and verified all respondent entries to ensure that only qualified participants are included in the study.
Executive Summary
Executive Summary

• Big data adoption reached 53 percent in 2017, up from 17 percent in 2015. Telecommunications and financial services are early adopters (pp. 19-24). Industry sentiment toward big data cooled slightly (pp. 68).
• Forty percent of non-users expect to adopt big data in the next two years, led by Asia-Pacific respondents (pp. 25-30).
• Among technologies and initiatives strategic to business intelligence, big data analytics ranks 20 out of 33 topical areas we study (p. 18).
• Data warehouse optimization remains the top big data use case. Customer/social analysis and predictive maintenance are the next most likely use cases. IoT big data momentum slowed (pp. 31-36).
• Spark is the leading big data infrastructure choice, followed by MapReduce and Yarn (pp. 37-42). Spark, MapReduce, and Yarn have the highest levels of vendor support (pp. 69-70).
• Spark SQL is the most popular means of big data access, followed closely by Hive and HDFS; respondents considered all three "important" (p. 43-48). Industry support is greatest for Hive / HiveQL, followed by Spark SQL. Industry big data access support is increasing (pp. 71-72).
• Among big data search mechanisms, Elasticsearch leads Apache Solr and Cloudera Search, though user demand is tepid (pp. 49-54). Big data search is gaining some industry investment but is not a high priority (pp. 73-74).
• The most popular big data analytics / machine learning technology is Spark MLib, followed by scikit-learn (pp. 55-60). Industry investment in machine learning is noticeably increasing, particularly for Spark MLib (pp. 75-76).
• Among big data distributions, Cloudera is the most popular, followed by Hortonworks, MAP/R, and Amazon EMR. Overall big data distribution sentiment declines somewhat in 2017 (pp. 61-66). Industry support is strong for big data distributions and increases for Hortonworks and Cloudera (pp. 77-78).
• Big data vendor ratings are shown on p. 80.
Study Demographics
Our 2017 Big Data Analytics Market Study is based on a cross-section of data that spans geographies, functions, organization size, and vertical industries. We believe that, unlike other industry research, this supports a more representative sample and better indicator of true market dynamics. We constructed cross-tab analyses using these demographics to identify and illustrate important industry trends.

Geography
North America, which includes the U.S., Canada, and Puerto Rico, represents 66 percent of respondents (fig. 1). EMEA accounts for the next largest group (24 percent), followed by Asia Pacific and Latin America.
Functions
IT (36 percent), the Business Intelligence Competency Center (BICC) (18 percent), and Executive Management (12 percent) are the largest groups represented in our big data analytics sample (fig. 2).

Examining trends and behavior by function helps us compare and contrast plans and priorities in different areas of organizations.

![Functions Represented](image-url)
**Vertical Industries**

Technology (14 percent), Healthcare (12 percent), and Financial Services (12 percent) are the most represented industries in our study, followed by Telecommunications, Education, and Consulting (fig. 3). We include responses from consultants—who often have greater interaction with initiatives and deeper industry knowledge than many customer counterparts. This also yields insight into the partner ecosystem for BI vendors.

![Figure 3 – Vertical industries represented](image-url)
Organization Size
Respondents to our big data analytics study reflect a mix of organizational sizes and structures (fig. 4). Small organizations of 1-100 employees represent 24 percent of the sample. Mid-sized organizations account for 22 percent, and the remaining 44 percent are large organizations with more than 1,000 employees.

Figure 4 – Organization sizes represented
Analysis and Trends
Analysis and Trends: Big Data Analytics

Importance of Big Data
Among technologies and initiatives considered strategic to business intelligence, big data analytics ranks 20 out of 33 topical areas we currently study (fig. 5). This ranking is identical to our 2016 Big Data Analytics Market Study. We should add that 2016 was a watershed year for increased big data adoption and perceived importance. While we continue to believe that big data interest varies widely from organization to organization, broader momentum clearly emerged in the last 24 months. Contextually, we also observe that big data still distantly trails the perceived importance of mainstream BI practices such as reporting, dashboards, and end-user self-service.

Figure 5 – Technologies and initiatives strategic to business intelligence
Big Data Adoption
Current users of big data (which we define as "systems that enable end-user access to and analysis of data contained and managed within the Hadoop ecosystem"), surpass 50 percent for the first time in 2017 (fig. 6). Another 36 percent of respondents say they may use big data in the future. Just 11 percent of respondents have "no plans to use big data at all" (an all-time low).

![Adoption of Big Data](image-url)
Across the three years of our comprehensive big data analytics study, we see a significant increase in uptake and a large drop in holdouts with no plans (fig. 7). Just 17 percent of respondents used big data in 2015, which appeared to debunk some of the extensive hype around big data two years ago. But that figure well more than doubled in 2016 to 41 percent. Now at 53 percent, big data has again grown 29 percent year over year. Correspondingly, the number of respondents with no plans fell by a factor of greater than two in 2016 and further still to just 11 percent in 2017.

**Adoption of Big Data 2015-2017**

![Adoption of Big Data 2015-2017](image)

*Figure 7 – Adoption of big data 2015-2017*
In our 2017 sample, North America (55 percent) narrowly leads EMEA (53 percent) in current adoption of big data by region (fig. 8). Asia-Pacific respondents report 44 current adoption and are most likely to say they "may use big data in the future." Slightly more North American and EMEA respondents (12 and 11 percent respectively) report "no plans to use big data at all."

Figure 8 – Adoption of big data by geography
By industry, current use of big data is greatest in telecommunications, where a robust 87 percent say they already adopted (fig. 9). Also impressive, 76 percent of Financial Services institutions are current users. By comparison, 61 percent of Technology industry respondents report current use; 20 percent have no plans, which is less intuitive given the industry’s first-mover reputation. Slightly less than 60 percent of Healthcare respondents are using big data and nearly all are open to future use. Higher Education is least likely to currently use big data (25 percent), but 67 percent are open to possible future use.

![Adoption of Big Data by Vertical Industry](image)

*Figure 9 – Adoption of big data by vertical industry*
In 2017, IT supplanted the BICC and R&D as the most likely current departmental user of big data (fig. 10). All departments, even Finance, are eyeing future use. This finding supports the notion that big data is moving from an experimental to practical pursuit in organizations. We note that Sales/Marketing is less a current user than most functions, though it is possible that needed services are coming from aforementioned IT and other resources.

Adoption of Big Data by Function

Figure 10 – Adoption of big data by function
Current adoption of big data is notably strongest (70 percent) within very large businesses and institutions with more than 5,000 employees (fig. 11). Only 6 percent of very large organizations have no plans for big data. Small (1-100), midsized (101-1,000) and some larger (1,001-5,000) organizations all report current adoption of 43 to 45 percent and are similarly open to future use of big data. We would expect that small organizations are most likely cloud users of big data services while large organizations will likely deploy on premises.

Adoption of Big Data by Organization Size

![Adoption of Big Data by Organization Size](image)

- Green: No. We have no plans to use big data at all
- Red: We may use big data in the future
- Blue: Yes. We use big data today

*Figure 11 – Adoption of big data by organization size*
Future Adoption of Big Data
Among organizations that have not yet adopted big data but have future plans (36 percent of the total sample) (fig.6, p. 19), 11 percent say they will adopt in the current calendar year (fig. 12). This horizon grows rapidly in 2018 when 28 percent plan to adopt. However, a large majority (61 percent) of organizations with future plans are deferring action to beyond 2018. Though we often find big data plans compartmentalized to projects or departments, future adoption will also hinge on enterprise plans for traditional and emergent technologies.

Future Adoption of Big Data

- Will adopt this year, 11.34%
- Will adopt next year, 27.84%
- Will adopt beyond next year, 60.82%

Figure 12 – Future adoption of big data
Compared to our inaugural 2015 study, "will adopt this year" (current year) plans for 2017 show a marked acceleration in near-term adoption (fig. 13). Put another way, current year adoption plans are almost three times higher in 2017 compared to 2015, while deferred adoption decreases. We also note that 12-month plans are more likely to be budgeted and usually more dependable than longer-term projections.

**Figure 13 – Future adoption of big data 2015-2017**
Regionally, among those that have not already adopted big data, North American organizations are most likely to adopt during the current year but also have the most plans deferred beyond next year (fig. 14). Asia-Pacific respondents are most likely to adopt next year. EMEA respondents have the fewest current year plans, and more than 60 percent are deferring adoption beyond next year.

![Future Adoption of Big Data by Geography](image-url)

*Figure 14 – Future adoption of big data by geography*
Among organizations not yet using big data, planned vertical adoption in 2017 is highest (about 50 percent) in Telecommunications (fig. 15). (Telecommunications is also the most penetrated sector so far) (fig. 9, p. 22.) Technology, followed by Financial Services organizations, are the next most likely to adopt this year. Healthcare and Education respondents are not predicting 2017 big data adoption.

**Future Adoption of Big Data by Vertical Industry**

![Future Adoption of Big Data by Vertical Industry](image)

*Figure 15 – Future adoption of big data by vertical industry*
Among non-users of big data, Executive Management reports the highest (20 percent) current-year adoption plans (fig. 16). R&D and IT are the next most likely to deploy big data this year, though at low rates of 14 percent and 12 percent respectively. Given that near-term delivery of technology is a leading indicator of uptake, these lower current year figures in R&D, IT, and the BICC (12 percent) may indicate some tempering of big data adoption in the near term.

**Future Adoption of Big Data by Function**

![Future Adoption of Big Data by Function](image)

*Figure 16 – Future adoption of big data by function*
Unlike other measures, current-year adoption by non-users at different-sized companies is mostly undistinguished by any particular trend (fig. 17). Between 10-13 percent of any size grouping from small to large will adopt big data in the current year. Next-year adoption grows somewhat with headcount, with the notable exceptions of very large organizations (>5,000 employees), where deferred adoption beyond next year is surprisingly highest at 68 percent.

Figure 17 – Future adoption of big data by organization size
Big Data Use Cases
The top big data use case in 2017 is data warehouse optimization, which is considered "critical" or "very important" to about 70 percent of respondents (fig. 18). "Customer/social analysis" and "predictive maintenance" (new for 2017), are the next most likely use cases that are, at minimum, "very important" to a majority of respondents. It is interesting to note that much-discussed IoT, a likely use case for big data, is the lowest priority among those we sampled.

Figure 18 – Big data use cases
Year over year, big data use cases retain their rankings (fig. 19). Sentiment toward data warehouse optimization and customer/social analysis changes little year over year. High mean scores (3.7 and 3.4 respectively) place their 2017 criticality at well above "important." IoT, the third most cited use case in 2015, falls to last place in 2017.

![Big Data Use Cases 2015-2017](image)

*Figure 19 – Big data use cases 2015-2017*
By region, Asia Pacific is most likely to prioritize all big data use cases, with the exception of clickstream analytics, which is more important to EMEA respondents (fig. 20). In all other cases, sentiment is second highest in EMEA and lowest in North America. Asia-Pacific respondents also report the only above 4.0 mean score in the case of data warehouse optimization, above a level of "very important."

**Big Data Use Cases by Geography**

![Big Data Use Cases by Geography](image)

*Figure 20 – Big data use cases by geography*
Big data use case priorities vary noticeably by industry (fig. 21). In 2017, the highest mean scores for data warehouse optimization are in Financial Services (4.2) and Healthcare (4.0). Elsewhere, data warehouse optimization is the second priority in Technology (behind customer/social analysis) and the least likely choice in Education, where customer/social is also most important. Expectedly, interest in fraud detection is highest in Financial Services and Telecommunications. Clickstream analysis is most popular in Financial Services.

**Figure 21 – Big data use cases by vertical industry**
In 2017, interest in big data use cases varies noticeably by function (fig. 22). In 2017, R&D, IT, and the BICC report the greatest interest in data warehouse optimization. R&D, along with IT, post the highest scores for predictive maintenance (new for 2017). As we might expect, Sales/Marketing and Executive Management are more interested in customer/social analysis than in any other use case. As we also found in 2016, Sales/Marketing has the greatest interest in clickstream analytics.

**Big Data Use Cases by Function**

![Bar chart showing big data use cases by function.](image-url)

*Figure 22 – Big data use cases by function*
Very large organizations (>5,000) expectedly have the greatest proportional interest in data warehouse optimization (fig. 23). Generally, we expect large organizations to be more conventional in their approach to big data use cases with an eye toward cost and operational efficiency, while smaller peers might be more opportunistic. In 2017, small organizations are most interested in customer/social analysis among organizations of any size.

**Figure 23 – Big data use cases by organization size**
Big Data Infrastructure
To gather baseline data on big data infrastructure awareness/adoptions, we assembled a list of relevant frameworks, databases, and other technologies in the Hadoop / open source orbits of interest (fig. 24). In our 2017 sample, Spark is the most preferred mechanism, followed by MapReduce and Yarn. Among the leaders, respondents consider Spark critical by more than 30 percent of respondents and, at minimum, "important" to 77 percent. MapReduce and Yarn are "critical" to more than 20 percent of respondents.

Figure 24 – Big data infrastructure
Across three years of study, Spark surpasses MapReduce as the preferred big data infrastructure (fig. 25). Preferences for Spark and associated applications/frameworks grew in 2017 and extend across all measures in this report even though MapReduce is also well penetrated. All other infrastructure choices except Yarn and Oozie lose some favor in 2017 compared to 2016.

Figure 25 – Big data infrastructure 2015-2017
All regions share the highest favor for Spark, particularly Asia Pacific and EMEA (fig. 26). Only the top two choices, Spark and MapReduce, earned mean scores of greater than 3.0, considered above "important. Thereafter, MapReduce and Oozie interest is highest in Asia Pacific, while North American respondents most prefer Yarn. Asia-Pacific respondents are generally more interested in big data infrastructure compared to North America.

Figure 26 – Big data infrastructure by geography
Big data infrastructure preferences vary by vertical industry (fig. 27). Financial Services and Telecommunications industry respondents noticeably have the most interest in Spark and Yarn (where Spark draws the only importance scores near or above 4.0, or "very important"). Oozie is also most popular in Financial Services and Telecommunications. Healthcare respondents are most likely to choose MapReduce as their top priority.

**Big Data Infrastructure by Vertical Industry**

![Graph showing big data infrastructure by vertical industry.](image-url)

*Figure 27 – Big data infrastructure by vertical industry*
Big data infrastructure preferences vary by function (fig. 28). R&D and the BICC (often contained within IT) are the strongest proponents of Spark. Outside of Finance, Spark also gets mean scores greater than 3.5 across all functions, between "important" and "very important." IT is slightly more likely than the BICC to be the leading proponent of MapReduce. R&D easily leads interest in Yarn, while IT is most likely to promote Oozie.

**Figure 28 – Big data infrastructure by function**
In 2017, organizations of different size have different priorities for big data infrastructure (fig. 29). As we might expect, Spark, MapReduce, Yarn, and Oozie are the preferred choice in very large organizations (>5,000 employees) compared to organizations of other sizes. It is interesting to note that mid-sized organizations (101-1,000 employees) are the next most interested in the same kinds of infrastructure. Thereafter, small organizations (1-100 employees) gather relative interest in lower-ranked technologies, notably Apache Drill.

**Figure 29 – Big data infrastructure by organization size**
**Big Data – Data Access**

We asked organizations which big data access methods they preferred and which are more/most important to them. This includes indirect access to Hadoop and other related engines. In our 2017 study, Spark SQL is the most cited and considered, at minimum, “important” to 73 percent of the sample (fig. 30). However, Hive and HDFS follow very closely and earned more "critical" scores (30 percent or more) than Spark (27 percent). We are not surprised to see a certain parochial audience dependent on subscription services like Amazon S3 in a cloud market that will be sorted out over time.

![Big Data - Data Access](image)

*Figure 30 – Big data – data access*
After strong growth in 2016, the top data access sources, Spark, Hive, and HDFS mostly held ground or grew slightly year over year (fig. 31). Only these three top-tier technologies receive scores greater than 3.0, above “important.” Mid-tier technologies including Hbase, Impala, MongoDB, Amazon Redshift, and Google BigQuery experience year-over-year declines.

Figure 31 – Big data - data access 2015-2017
Big data access preferences vary by region (fig. 32). Asia-Pacific respondents have the strongest response to most upper-tier technologies and obviously stronger responses to subscription services including Amazon S3, Redshift, and Dynamo DB. Asia-Pacific sentiment is more muted for Google BigQuery. EMEA respondents conversely lead interest in Google BigQuery as well as Kudu. North American sentiment trails that in other regions, with the top choices Spark, Hive, and HDFS.

![Big Data - Data Access by Geography](image_url)

*Figure 32 – Big data – data access by geography*
By vertical industry, Financial Services, Telecommunications, and Healthcare are the most aligned around Spark SQL, Hive, and HDFS for data access (fig. 33). The same two industries lead interest in several lesser priorities. Impala and Cassandra are more popular in Financial Services than in other industries. Technology industry organizations show relatively high interest in service-based Amazon S3, Redshift, and Dynamo DB, along with Google BigQuery.

Figure 33 – Big data – data access by vertical industry
Departmental interest in data access varies by function (fig. 34). BICC and IT respondents report high scores for all top data access types, particularly Hive, Spark, and Amazon S3. R&D also favors the top three choices, particularly the baseline HDFS framework. By a slight margin, the highest overall interest in Spark, Impala, and Amazon Redshift comes from Sales/Marketing. Executives are drawn to several areas that include service-based offerings, particularly Amazon S3.

![Big Data - Data Access by Function](image)

*Figure 34 - Big data - data access by function*
Very large organizations are most drawn to big data access through Spark, Hive, and HDFS (fig. 35). However, mid-sized organizations (101-1,000) are next most drawn to the same top choices compared to some larger organizations (1,001-5,000 employees). Small organizations (1-100 employees) are proportionately most drawn to lower-risk service-based Amazon S3, Redshift, and Google BigQuery, along with MongoDB and lower priorities such as Kudu.

**Big Data - Data Access by Organization Size**

![Big Data - Data Access by Organization Size](image-url)

*Figure 35 – Big data - data access by organization size*
**Big Data Search**

We asked respondents to rank interest in big data search facilities, which in Hadoop include indexing and natural language textual search (fig. 36). In our 2017 sample, Elasticsearch resonates most strongly, followed by Apache Solr and Cloudera Search. While Elasticsearch holds a visible lead, there is no dominant choice in big data search, and "very important" to "critical" importance applies to only 27 to 35 percent of respondents for any tool.

![Big Data Search](image)

*Figure 36 – Big data search*
Across three years of study data, we saw a small reversal of fortunes among big data search options (fig. 37). Elasticsearch moved past Cloudera and Apache Solr in 2016 and has additional interest in 2017 compared to the other two choices. 2017 sentiment toward all big data search choices remains between 2.6 and 2.9, well above "somewhat important" but below "important."

**Big Data Search 2015-2017**

![Big Data Search 2015-2017](image)

*Figure 37 – Big data search 2015-2017*
All three geographic regions favor Elasticsearch over other choices in 2017 (fig. 38). Only EMEA respondents score all three big data search choices above 3.0 or "important." Where EMEA and North American respondents make Apache Solr their second choice, Asia-Pacific respondents choose Cloudera Search over open source Apache Solr.

Figure 38 – Big data search by geography
We see some divergence from overall results in big data search preference by industry (fig. 39). All industries with the exception of Financial Services pick Elasticsearch as their top choice. The highest rankings for Cloudera Search are seen in Financial Services, followed by Telecommunications. The highest rankings for Apache Solr come from Financial Services and Technology. Only Financial Services ranks all three technologies higher than 3.0 or "important."

![Big Data Search by Vertical Industry](image_url)

*Figure 39 – Big data search by vertical industry*
All functions with the exception of Finance choose Elasticsearch as their top big data search priority (fig. 40). As a technology still in early stage adoption, it's worthy to note the considerable R&D activity in Elasticsearch compared to other choices. Likewise, Executive Management, likely to keep an eye on the technology horizon, along with most other functions, shows a bit more interest in Elasticsearch than other choices.

**Figure 40 – Big data search by function**
Organizations of different size, with the exception of small organizations (1-100 employees), choose Elasticsearch as their first big data search preference (fig. 41). Small organizations slightly prefer open source Apache Solr over other choices. Very large (>5,000 employees) and mid-sized organizations (101-1,000 employees) show broader interest than large organizations with 1,001 to 5,000 employees.

![Big Data Search by Organization Size](image)

*Figure 41 – Big data search capabilities by organization size*
Big Data Analytics / Machine-Learning Technologies

We asked respondents to rank their interest in a variety of big data analytics and machine-learning technologies (fig. 42). The leader, Spark MLib (here and throughout this category), is considered, at minimum, “important” by close to than 60 percent of respondents and ranks well ahead of most competitors. New for our 2017 report, however, scikit-learn weighed in at second place ahead of more traditional technologies Rhipe and Mahout.

Figure 42 – Big data analytics / machine learning
Year-over-year interest in big data analytics and machine learning jumped in 2016 but declines somewhat in 2017, though perhaps not to a statistically important level (fig. 43). As mentioned in on the previous page (fig. 42), scikit-learn debuts in our survey as second choice; but only Spark MLib holds scores close to 3.0 or "important" in 2016-2017.

Figure 43 – Big data analytics / machine learning 2015-2017
All geographic regions sampled choose Spark MLib over all other technologies as their top machine-learning capability in 2017 (fig. 44). Asia-Pacific respondents particularly prefer Spark MLib and also make Mahout and Rhipe their second and third priorities. EMEA show the strongest response for scikit-learn, and it also ranks as the second most important in North America.

![Big Data Analytics / Machine Learning by Geography](http://www.dresneradvisory.com)

Figure 44 – Big data analytics / machine learning by geography
In 2017, all but one industry chooses Spark MLib as their top big data analytics / machine learning mechanism (fig. 45). The exception, Financial Services, shows a slight preference for scikit-learn. Overall, Spark MLib and scikit-learn are the top two choices in all industries except Technology, which makes Mahout and Rhippe its second and third choices respectively.

**Figure 45 – Big data analytics / machine learning by vertical industry**
By function, Spark MLib is the standout category leader across all organizational roles with the exception of Finance (fig. 46). Spark MLib and scikit-learn are the top two choices in BICC, Sales/Marketing, R&D, and IT, though IT is less engaged with big data analytics and machine learning compared to all other functions except Finance. R&D remains most engaged with the top two choices, indicating a horizon for future use.

**Big Data Analytics / Machine Learning by Function**

![Bar chart showing usage of different tools by function](http://www.dresneradvisory.com/figures/46.png)

*Figure 46 – Big data analytics / machine learning by function*
Organizations of all sizes prefer Spark MLib over all other big data analytics / machine-learning options, though this effect is not entirely linear (fig. 47). Very large organizations (>5,000 employees) and other large organizations with 1,001 to 5,000 employees are likely to refer Spark MLib and scikit-learn over other choices. Small organizations (1-100 employees) and mid-sized organizations (101-1,000 employees) have strong MLib sentiment but more mixed attitudes toward Rhipe, scikit-learn, and Mahout.

**Figure 47 – Big data analytics / machine learning by organization size**
Big Data Distributions
We asked respondents to rank the most important big data distributions by order of importance, and their responses show fairly clear hierarchy (fig. 48). In 2017, Cloudera led in measures of "critical" and was the strongest overall performer, followed by Hortonworks, MAP/R, and Amazon EMR. Only Cloudera, Hortonworks, and MAP/R are all seen as, at minimum, "important" to majorities of 54 percent to 62 percent of respondents.

Figure 48 – Big data distributions
Following a jump between 2015 and 2016, 2017 interest in all big data distributions measured over time decreases somewhat year over year while retaining their rankings (fig. 49). Cloudera, Hortonworks, MAP/R, and Amazon EMR remain the top four choices by order while newly covered Microsoft HD insights, Google Dataproc, and IBM BigInsights round out the rankings.

Figure 49 – Big data distributions 2015-2017
By region, EMEA respondents show the most interest overall in Cloudera, which was the second choice in most other regions (fig. 50). By order of priority, rankings are mostly consistent across all geographies in 2017. Asia Pacific is again the leader, almost across the board, on all distribution interest.

**Big Data Distributions by Geography**

- Cloudera
- Hortonworks
- MAP/R
- Amazon EMR
- Microsoft HD Insights
- Google Dataproc
- IBM BigInsights

![Figure 50 – Big data distributions by geography](image-url)
By vertical industry, Financial Services, Technology, and Telecommunications reported standout interest in Cloudera versus all other big data distributions (fig. 51). After this finding, industry results are more equivocal. Hortonworks is the second choice in Technology and Telecommunications but trails MAP/R in Financial Services. Healthcare marginally chooses Hortonworks and MAP/R over other distributions. Education is most interested in MAP/R but, overall, is less enthusiastic than other industries.

**Figure 51 – Big data distributions by vertical industry**
By function, Cloudera is a top choice among big data distributions across all roles (fig. 52). In our 2017 sample, Hortonworks is most popular among BICC respondents and the second choice for IT and Executive Management respondents but a distant last among Sales/Marketing respondents. Sales/Marketing report the greatest interest in MAP/R. Amazon EMR is most interesting to R&D audiences and performs well with Executive Management and other audiences.

![Big Data Distributions by Function](image)

**Figure 52 – Big data distributions by function**
To varying extents, small to very large organizations have different preferences toward big data distributions (fig. 53). Very large and mid-sized organizations (101-1,000 employees) most prefer Cloudera and Hortonworks. As we might expect, cloud-based Amazon EMR and Google Dataproc distributions appeal most strongly to small organizations for simple and inexpensive startup projects that also demonstrate ability to scale. MAP/R and IBM BigInsights are also most interesting to small and mid-sized organizations.

**Figure 53 – Big data distributions by organization size**
Industry and Vendor Analysis
Industry and Vendor Analysis
In our third comprehensive big data study, we reached out to the vendor community with questions about their capabilities and investment plans for technologies in big data analytics. Compared to 2016, industry sentiment appears to have bifurcated somewhat (fig. 54). A majority (51 percent) of vendors say big data is "critical," up from 47 percent in 2016. But fewer vendors (28 percent) currently say big data is "very important," compared to 44 percent in 2016. Overall, we might say industry urgency cools somewhat in 2017, perhaps because most industry vendors now have products in the marketplace. Along with ongoing increases in user adoption (fig. 7, p. 20), we might also observe that both supply and demand for big data technologies are coming into balance.

![Industry Importance of Big Data 2015-2017](image-url)

*Figure 54 – Industry importance of big data 2015-2017*
Among big data infrastructure options in the Hadoop ecosystem, Spark and MapReduce have the highest levels of vendor support (fig. 55). After Yarn and Tez, current support for big data infrastructure diminishes rapidly: all trailing technologies likely will have less than 50 percent and as little as 19 percent support 12 months from today. That said, those customers that rely on more esoteric big data infrastructure going forward are very likely to be supported.

**Industry Support for Big Data Infrastructure**

![Industry Support for Big Data Infrastructure](image)

*Figure 55 – Industry support for big data infrastructure*
In 2017, industry support for Spark surpasses MapReduce as the top infrastructure choice as the latter lost momentum (fig. 56). To a lesser extent than Spark, Yarn and Tez also show strong sustained support momentum over time. Among lesser chosen infrastructure types, Knox Gateway, Mesos, and Atlas show positive momentum. Particularly among top choices, industry support is mostly aligned with user sentiment (fig. 25, p. 38). We continue to expect that proprietary vendor support of open source big data projects will be opportunistic and customer driven.

Figure 56 – Industry support for big data infrastructure 2015-2017
Existing industry support for access to big data sources is greatest for Hive / HiveQL (94 percent), followed by Spark SQL (91 percent) (fig. 57). These two choices are in line with top user preferences for data access (fig. 30, p. 43), though less aligned among lesser choices. Amazon Redshift, for example, is well ahead of user rankings by requirement, while Amazon S3 support trails user rankings.

Figure 57 – Industry support for access to big data sources
Year-over-year industry support for data access increases for all big data sources we measured (fig. 58). The rapid increase of Spark support reflects industry responsiveness to user demand (fig. 31, p. 44). Google BigQuery support grew by more than 50 percent year over year but apparently not in response to accelerated user demand. Respondents added several other data access sources for 2017, which have no current trending data.

Figure 58 – Industry support for access to big data sources 2015-2017
Industry support for big data search remains tepid in 2017, though some investment plans are in place (fig. 59). Just 35 percent of vendors currently support Elasticsearch compared to 33 percent for Apache Solr and 26 percent for Cloudera Search. Despite user preferences toward Cloudera distributions (fig. 48, p. 61), considerably more industry investment is forecast for Elasticsearch and, to a lesser extent, Apache Solr. User sentiment for big data search remains weak, as shown in fig. 36, p. 49.

**Figure 59 – Industry support for big data search**
Year-over-year industry support for big data search increases in 2017 for all the choices we surveyed (fig. 60). As implied in the previous chart (fig. 59, p. 73), momentum grew most strongly for Elasticsearch. Though it trails the category, Cloudera Search was the next largest gainer in industry support from 2016 to 2017. User interest in all three big data search products flattens year over year (fig. 37, p. 50).

**Industry Support for Big Data Search 2015-2017**

![Industry Support for Big Data Search 2015-2017](image)

*Figure 60 – Industry support for big data search 2015-2017*
Industry support and investment plans for big data analytics / machine learning is more striking than in other categories. Current support is highest for Spark MLib (46 percent), which will grow by 60 percent in the next 12 months to 72 percent and is projected to reach almost 80 percent industry support in future time frames (fig. 61). Scikit-learn is the next most supported (31 percent) and should surpass 50 percent in future time frames. After Spark MLib, 12-month support grows fastest for lesser-penetrated Mahout, which is expected to surpass Rhipe support in 2018. (Spark MLib, scikit-learn, and Rhipe are the top user machine-learning choices in 2017 (fig. 42, p. 55).

![Industry Support for Big Data Analytics / Machine Learning](image_url)
Year-over-year industry support for big data analytics / machine learning is easily highest and actually doubles for Spark MLib from 23 percent to 46 percent (fig. 62). Scikit-learn, new for our 2017 study, is currently the next most supported. Elsewhere, Rhipe support surpassed Mahout in 2017, though vendor projections as shown on the previous page (fig. 61) will return them to near-parity support.

**Industry Support for Big Data Analytics / Machine Learning 2015-2017**

![Industry Support for Big Data Analytics / Machine Learning 2015-2017](image-url)

*Figure 62 – Industry support for big data analytics / machine learning 2015-2017*
Industry support is relatively strong for big data distributions, led by Hortonworks (86 percent), closely followed by Cloudera (83 percent) (fig. 63). Map/R support is next strongest with investment plans to reach near 80 percent support in future time frames. More significant future plans are forecast for currently trailing service-based distributions, particularly Microsoft HD Insights.

Figure 63 – Industry support for big data (Hadoop) distributions
Industry support/investments in Hortonworks and Cloudera big data distributions grow year over year in 2017, while support for MAP/R and Amazon decline or are mostly flat (fig. 64). Microsoft HD Insights, IBM BigInsights, and Google Dataproc are newly tracked in 2017 but are not yet driving most big data distribution momentum.

![Industry Support for Big Data (Hadoop) Distributions 2015-2017](image)

Figure 64 – Industry support for big data (Hadoop) distributions 2015-2017
Big Data Analytics Vendor Ratings

In rating vendors for big data analytics, we examined levels of functionality in five categories: infrastructure, data access, search, machine learning, and supported distributions (fig. 65). Criteria were weighted based on user responses/priorities. Top-rated vendors include Zoomdata (1st), Arcadia Data (2nd), MicroStrategy (3rd), RapidMiner (3rd), Hitachi Vantara (4th), and TIBCO (5th).

![Big Data Analytics Vendor Ratings](http://www.dresneradvisory.com)

Figure 65 – Big data analytics vendor ratings
Vendor Ratings Detail
Often, groupings of scores are tight with small percentages separating one vendor from another. To help readers understand which vendors offer key capabilities, we identified top-rated vendors by categories of big data analytics functionality, specifically “infrastructure,” “data access,” “search,” “machine learning,” and “distribution support.”

Top Vendors for Big Data Infrastructure
With scores ranging from 5.5 to 7 (maximum score of 7.25), top vendors for big data infrastructure include Arcadia Data, Hitachi Vantara, RapidMiner, Zoomdata, SAS, TIBCO, and MicroStrategy.

Top Vendors for Big Data / Data Access
With scores ranging from 11.25 to 14 (maximum score of 14.25), top vendors for big data / data access include Infor/Birst, Logi Analytics, RapidMiner, Sisense, Zoomdata, MicroStrategy, TIBCO, Arcadia Data, and OpenText.

Top Vendors for Big Data Search
With a score of 2.75 (maximum score of 2.75), top vendors for big data search include Arcadia Data, Domo, MicroStrategy, and Zoomdata.

Top Vendors for Big Data Machine Learning
With scores ranging from 1.75 to 2 (maximum score of 2.75), top vendors for big data search include RapidMiner, Hitachi Vantara, OpenText, SAS, and TIBCO.

Top Vendors for Big Data Distribution Support
With scores ranging from 4.75 to 5.5 (maximum score of 5.5), top vendors for big data search include Infor/Birst, Zoomdata, Arcadia Data, Datameer, MicroStrategy, Hitachi Vantara, SAS, and Sisense.
**Glossary**

**Alluxio** (formerly Tachyon) is a memory-centric distributed storage system enabling reliable data sharing at memory-speed across cluster frameworks.

*Source: alluxio.org*

**Atlas** is designed to exchange metadata with other tools and processes within and outside of the Hadoop stack, thereby enabling platform-agnostic governance controls that effectively address compliance requirements.

*Source: Apache Software Foundation*

**BigQuery** is a RESTful web service that enables interactive analysis of massively large datasets working in conjunction with Google Storage. It is an Infrastructure as a Service (IaaS) service that may be used complementarily with MapReduce.

**Elasticsearch** is a search server based on Lucene. It provides a distributed, multitenant-capable full-text search engine with an HTTP web interface and schema-free JSON documents. Elasticsearch is developed in Java and is released as open source under the terms of the Apache License. Elasticsearch is the second most popular enterprise search engine after Apache Solr.*

**HAWQ** is a parallel SQL query engine that combines the key technological advantages of the industry-leading Pivotal Analytic Database with the scalability and convenience of Hadoop. HAWQ reads data from and writes data to HDFS natively. HAWQ delivers industry-leading performance and linear scalability. It provides users the tools to confidently and successfully interact with petabyte range data sets. HAWQ provides users with a complete, standards-compliant SQL interface.

*Source: Pivotal*

**HBase** is an open source, non-relational, distributed database modeled after Google's BigTable and is written in Java. It is developed as part of Apache Software Foundation's Apache Hadoop project and runs on top of HDFS (Hadoop Distributed File System), providing BigTable-like capabilities for Hadoop.

The Hadoop distributed file system (**HDFS**) is a distributed, scalable, and portable file system written in Java for the Hadoop framework.

The **Apache Hive™** data warehouse software facilitates querying and managing large datasets residing in distributed storage. Hive provides a mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL. At the same time this language also allows traditional MapReduce programmers to plug in their custom mappers and reducers when it is inconvenient or inefficient to express this logic in HiveQL.
Source: Apache Software Foundation

The **Apache Knox Gateway** is a REST API Gateway for interacting with Apache Hadoop clusters. The Knox Gateway provides a single access point for all REST interactions with Apache Hadoop clusters.

Source: Apache Software Foundation

**Impala** is an open source, native analytic database for Apache Hadoop. Impala is shipped by Cloudera, MapR, Oracle, and Amazon.

Source: Cloudera

**Mahout** is a project of the Apache Software Foundation to produce free implementations of distributed or otherwise scalable machine learning algorithms focused primarily in the areas of collaborative filtering, clustering and classification. Many of the implementations use the Apache Hadoop platform. Mahout also provides Java libraries for common math operations (focused on linear algebra and statistics) and primitive Java collections.

Source: Apache Software Foundation

**MapReduce** is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster. Conceptually similar approaches have been very well known since 1995 with the Message Passing Interface standard having reduce and scatter operations.

Apache **Mesos** is an open source cluster manager that was developed at the University of California, Berkeley. It "provides efficient resource isolation and sharing across distributed applications, or frameworks". The software enables resource sharing in a fine-grained manner, improving cluster utilization.

**MLlib** is Spark’s scalable machine-learning library consisting of common learning algorithms and utilities, including classification, regression, clustering, collaborative filtering, dimensionality reduction, as well as underlying optimization primitives.

Source: Apache Software Foundation

**MongoDB** is a cross-platform document-oriented database. Classified as a NoSQL database, MongoDB eschews the traditional table-based relational database structure in favor of JSON-like documents with dynamic schemas (MongoDB calls the format BSON), making the integration of data in certain types of applications easier and faster. Released under a combination of the GNU Affero General Public License and the Apache License, MongoDB is free and open source software.
**Myrrix**, offers a “complete, real-time, scalable clustering and recommender system.” The solution is built on top of the Apache Mahout machine-learning project.

*Source: Cloudera*

**Oozie** is a workflow scheduler system to manage Hadoop jobs. It is a server-based Workflow Engine specialized in running workflow jobs with actions that run Hadoop MapReduce and Pig jobs. Oozie is implemented as a Java Web application that runs in a Java servlet container.

*Source: Cloudera*

**Oryx** is built on Apache Spark and Apache Kafka, with specialization for real-time large scale machine learning. It is a framework for building applications but also includes packaged, end-to-end applications for collaborative filtering, classification, regression, and clustering.

*Source: Cloudera*

**RHIPE** integrates the R statistical environment with the Hadoop framework. RHIPE allows R users to compute on terabyte-sized data sets a cluster using the MapReduce framework, thus offering the best of both worlds to users seeking to leverage the strength of R and Hadoop. People with very large data sets stored in the Hadoop Distributed File System can now easily process the data on hundreds or even thousands of nodes in parallel, using only the R language.

*Source: Revolution Analytics*

**Cloudera Search** is one of Cloudera's near-real-time access products. Cloudera Search enables non-technical users to search and explore data stored in or ingested into Hadoop and HBase. Users do not need SQL or programming skills to use Cloudera Search because it provides a simple, full-text interface for searching.

*Source: Cloudera*

**Solr** is an open source enterprise search platform, written in Java, from the Apache Lucene project. Its major features include full-text search, hit highlighting, faceted search, real-time indexing, dynamic clustering, database integration, NoSQL features and rich document (e.g., Word, PDF) handling. Providing distributed search and index replication, Solr is designed for scalability and fault tolerance. Solr is the most popular enterprise search engine.

*Source: Cloudera*

**Apache Spark** is an open source cluster computing framework originally developed in the AMPLab at University of California, Berkeley but was later donated to the Apache Software Foundation where it remains today. In contrast to Hadoop's two-stage disk-based MapReduce paradigm, Spark's multi-stage in-memory primitives provides performance up to 100 times faster for certain applications. By allowing user programs to load data into a cluster's memory and query it repeatedly, Spark is well suited to machine-learning algorithms.
Spark SQL is a component on top of Spark Core that introduces a new data abstraction called DataFrames, which provides support for structured and semi-structured data. Spark SQL provides a domain-specific language to manipulate DataFrames in Scala, Java, or Python. It also provides SQL language support with command-line interfaces and ODBC/JDBC server.

Apache™ Tez is an extensible framework for building high-performance batch and interactive data-processing applications, coordinated by YARN in Apache Hadoop. Tez improves the MapReduce paradigm by dramatically improving its speed while maintaining MapReduce’s ability to scale to petabytes of data. Important Hadoop ecosystem projects like Apache Hive and Apache Pig use Apache Tez, as do a growing number of third-party data-access applications developed for the broader Hadoop ecosystem.

Source: Apache Software Foundation

YARN is one of the key features in the second-generation Hadoop 2 version of the Apache Software Foundation’s open source distributed processing framework. Originally described by Apache as a redesigned resource manager, YARN is now characterized as a large-scale, distributed operating system for big data applications.

*All sources Wikipedia unless otherwise noted*
Other Dresner Advisory Services Research Reports

- “Flagship” Wisdom of Crowds® Analytical Data Infrastructure Market Study
- “Flagship” Wisdom of Crowds® Business Intelligence Market Study
- “Flagship” Wisdom of Crowds® Enterprise Planning market Study
- Advanced and Predictive Analytics
- Business Intelligence Competency Center
- Cloud Computing and Business Intelligence
- Collective Insights®
- Embedded Business Intelligence
- End User Data Preparation
- IoT Intelligence®
- Location Intelligence
- Small and Mid-Sized Enterprise BI
Appendix: Big Data Analytics Study Survey Instrument

Please provide your contact information below:

Name*: _________________________________________________

Company Name: ____________________________________________

Address 1: ________________________________________________

Address 2: ________________________________________________

City: _____________________________________________________

State: ____________________________________________________

Zip: ______________________________________________________

Country: __________________________________________________

Email Address*: __________________________________________

Phone Number: ____________________________________________

Major Geography

( ) Asia/Pacific

( ) Europe, Middle East and Africa

( ) Latin America

( ) North America

What is your current title?

______________________________________________________
What function are you a part of?

( ) Business intelligence competency center

( ) Executive management

( ) Finance

( ) Information Technology (IT)

( ) Manufacturing

( ) Marketing

( ) Project/program management office

( ) Sales

( ) Research and development (R&D)

( ) Other - Write In: _______________________________ ________________

Please select an industry

( ) Advertising

( ) Aerospace

( ) Agriculture

( ) Apparel and accessories

( ) Automotive

( ) Aviation

( ) Biotechnology

( ) Broadcasting

( ) Business services

( ) Chemical

( ) Construction
Sports
State and local government
Technology
Telecommunications
Transportation
Utilities
Other - Write In: ____________________________

How many employees does your company employ worldwide?

1 - 100
101 - 1000
1001 - 5000
More than 5000

Do you use or intend to use big data technology/architecture within your organization?*

Yes. We use big data today
No. We have no plans to use big data at all
We may use big data in the future

What product(s) does your organization use with big data for BI/analytics?

____________________________________
____________________________________
____________________________________

How satisfied are you with your vendor and product for big data analytics?

Extremely satisfied
Somewhat satisfied
Somewhat unsatisfied
( ) Unsatisfied

What are your plans for Big data (Hadoop) Analytics in the Future?

( ) Will adopt this year
( ) Will adopt next year
( ) Will adopt beyond next year

What use cases are most important for Big data (Hadoop) in your organization?

<table>
<thead>
<tr>
<th>Use Case</th>
<th>Critical</th>
<th>Very important</th>
<th>Important</th>
<th>Somewhat important</th>
<th>Not important</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data warehouse optimization</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Customer/social analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internet of things</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fraud detection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clickstream analytics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predictive maintenance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Please indicate the importance of the following Big data infrastructure components

<table>
<thead>
<tr>
<th>Component</th>
<th>Critical</th>
<th>Very important</th>
<th>Important</th>
<th>Somewhat important</th>
<th>Not important</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tacyon</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
</tr>
<tr>
<td>Mesos</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
</tr>
<tr>
<td>Spark</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
</tr>
<tr>
<td>MapReduce</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
</tr>
<tr>
<td>Oozie</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
</tr>
<tr>
<td>Yarn</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
</tr>
<tr>
<td>Tez</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
</tr>
<tr>
<td>Atlas</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
</tr>
<tr>
<td>Knox Gateway</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
</tr>
<tr>
<td>Apache Drill</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
</tr>
</tbody>
</table>
Please indicate the importance of the following Big Data - data access capabilities

<table>
<thead>
<tr>
<th></th>
<th>Critical</th>
<th>Very important</th>
<th>Important</th>
<th>Somewhat important</th>
<th>Not important</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google BigQuery</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>HBase</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>HDFS</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>Hive/HiveQL</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>Impala</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>MongoDB</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>Pivotal HAWQ</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>Amazon Redshift</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>Spark SQL</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>CouchDB</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>Cassandra</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>Amazon S3</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>Presto</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>Kudu</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>Amazon DynamoDB</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
</tbody>
</table>
Please indicate the importance of the following Big data search capabilities

<table>
<thead>
<tr>
<th></th>
<th>Critical</th>
<th>Very important</th>
<th>Important</th>
<th>Somewhat important</th>
<th>Not important</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloudera Search</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>Apache Solr</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>Elasticsearch</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
</tbody>
</table>

Please indicate the importance of the following Big data analytical/machine learning components

<table>
<thead>
<tr>
<th></th>
<th>Critical</th>
<th>Very important</th>
<th>Important</th>
<th>Somewhat important</th>
<th>Not important</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mahout</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>Rhipe (R)</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>Oryx</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>Myrrix</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>Spark MLib</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>scikit-learn</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
</tbody>
</table>
Please indicate the importance of the following Big data (Hadoop) distributions

<table>
<thead>
<tr>
<th></th>
<th>Critical</th>
<th>Very important</th>
<th>Important</th>
<th>Somewhat important</th>
<th>Not important</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloudera</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
</tr>
<tr>
<td>Hortonworks</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
</tr>
<tr>
<td>MAP/R</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
</tr>
<tr>
<td>Amazon EMR</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
</tr>
<tr>
<td>Google Dataproc</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
</tr>
<tr>
<td>Microsoft HD Insights</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
</tr>
<tr>
<td>IBM BigInsights</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
<td>( )</td>
</tr>
</tbody>
</table>