Do You Hadoop?
A Survey of Big Data Practitioners

October 29, 2013

Bradley Graham
M. R. Rangaswami

SandHill Group

Executive Summary

- Expanding implementation of advanced analytics (e.g., risk, propensity, optimization)
  - Nearly one in four (24.4%) respondents stated that advanced analytics will be their #1 Hadoop-based initiative in the next 12-18 months. This almost threefold increase over the 8.9% currently developing advanced analytics is emblematic of the larger shift towards initiatives that can have a transformational impact on the organization.
- Increasing use of new data types
  - Over the coming 12-18 months, 20% more companies plan to incorporate new data types (such as streaming data and geographic data) into their Hadoop environment in support of new advanced analytics.
- Moderating emphasis on basic analytics
  - In the same timeframe, 32% of the respondents expressed a move away from basic analytics — such as statistics, patterns and search — as they ramp up their use of advanced analytics to enrich their business perspective and improve results.

Moving from Talk/Discussion to Action

- 44.4% of those surveyed indicated their company is currently in the exploration and education phase, building the experience and competencies needed for a successful Hadoop/Big Data initiative.
  - Top uses of Hadoop/Big Data within this segment are:

Companies in Exploration and Education Phase	All Companies
Basic analytics (60.0%)	Basic analytics (58.5%)
Business intelligence (46.7%)	Business intelligence (48.1%)
Archive more data (35.0%) Data preparation (35.0%)	Data preparation (45.9%)

- 16.3% are currently conducting a formal proof of concept (POC).
- 11.1% are in the process of developing their first Hadoop-based solution.

Satisfaction Levels Highlight the Challenges Ahead

- There is a significant gap between those who believe the results of their Hadoop/Big Data initiatives were better/less than expected:
  - 35.6% of companies indicated results were somewhat or significantly less than expected.
  - Only 11.1% of respondents said results achieved were either somewhat or significantly better than expected.
  - The three most common challenges associated with implementing Hadoop/Big Data initiatives are:

Most Commonly Reported #1 Hadoop-Related Challenges	Top Hadoop-Related Challenges
Knowledge and experience (46.7%)	Knowledge and experience (65.2%)
Skills availability (20.7%)	Skills availability (52.6%)
Development effort (6.7%)	Development effort (40.7%)

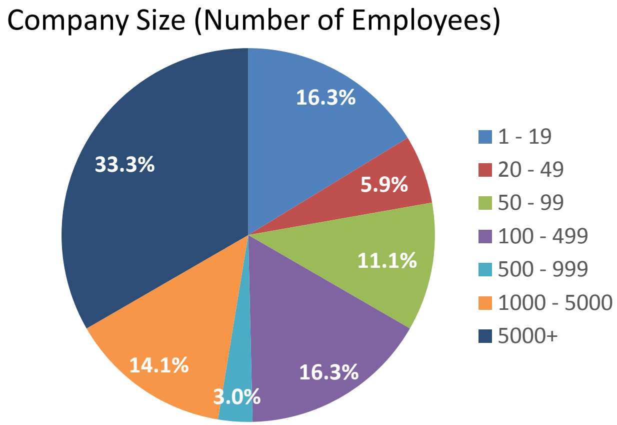
Research Objectives

- Clarify where companies are in the process of implementing Hadoop-based solutions
- Illuminate how the use of Hadoop will change over the next 12-18 months
- Determine the level of satisfaction with the experience and results to date
- Identify key pain points and barriers in the adoption process

Methodology

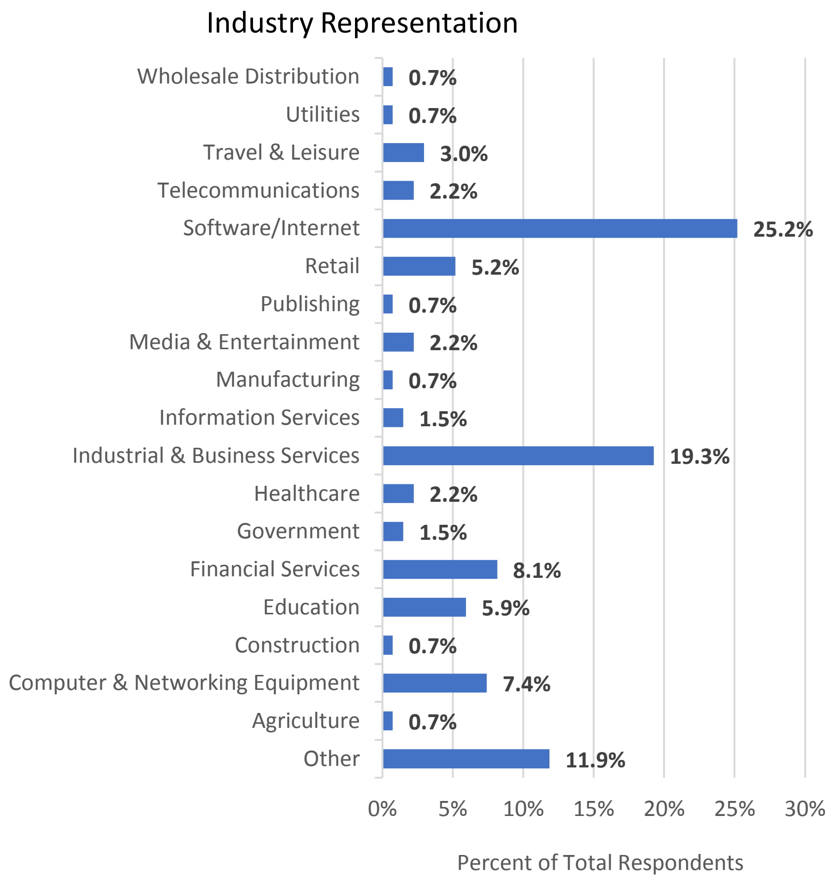
- Data was collected in early October 2103 by Sand Hill Group using an online survey designed specifically to capture the experience and plans of companies currently working on Hadoop-based projects.
- Invitations to participate in the research were sent via email by Sand Hill Group and other partner firms, Carpe Datum Rx and The Hive.
- The invitation also was posted to various industry and professional websites including LinkedIn and Twitter and was sent to several lists and related discussion groups.
- A total of 135 valid surveys were obtained and are the basis of the analyses in this report.
- The survey included personal demographic questions but did not require all of the information to be provided. In instances where information was not provided, it was categorized as “Other” or “Did Not Provide,” whichever was most applicable.

Demographics – Company Size



- The size metric used to segment the respondent population was the number of employees. For analytic purposes the seven size categories offered in the survey and shown in the adjacent chart were consolidated to three:
  - Small (33.3%): 1 - 99 employees
  - Medium (19.3%): 100 - 999 employees
  - Large (47.4%): 1000+ employees
- The diverse mix of company sizes offers insight into the differences that the scale of the business has as to when and how Hadoop/Big Data initiatives are implemented as well as the satisfaction with the results achieved to date.
- Respondents represent numerous brand-name companies across industries such as agriculture, computer and networking systems, consulting, financial services, retail and software/Internet.

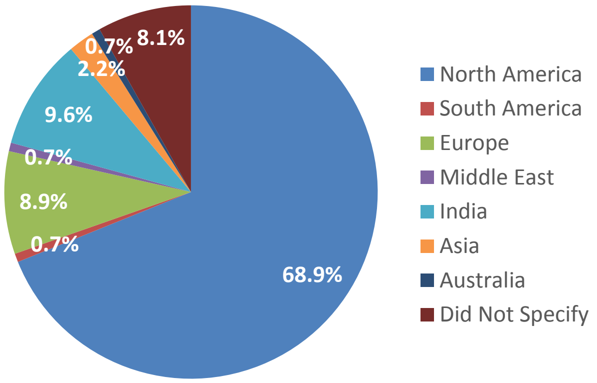
Demographics - Industry



- Respondents represent a wide variety of industries including many not traditionally viewed as information technology leaders or data centric.
- Technology-related industries account for slightly more than half (51.9%) of the companies included in the study.
- Industrial and business services encompasses companies offering:
  - IT professional services
  - Management consulting
  - Other non-technology related services (e.g., environmental services)
- Depending on the respondent’s position in the company, submissions from consulting firms reflect the use of Hadoop and Big Data for either internal projects or client engagements.
- Likewise for technology companies, submissions may reflect internal use, product initiatives or client-related work.

Demographics – Geographical Location

Geographical Representation

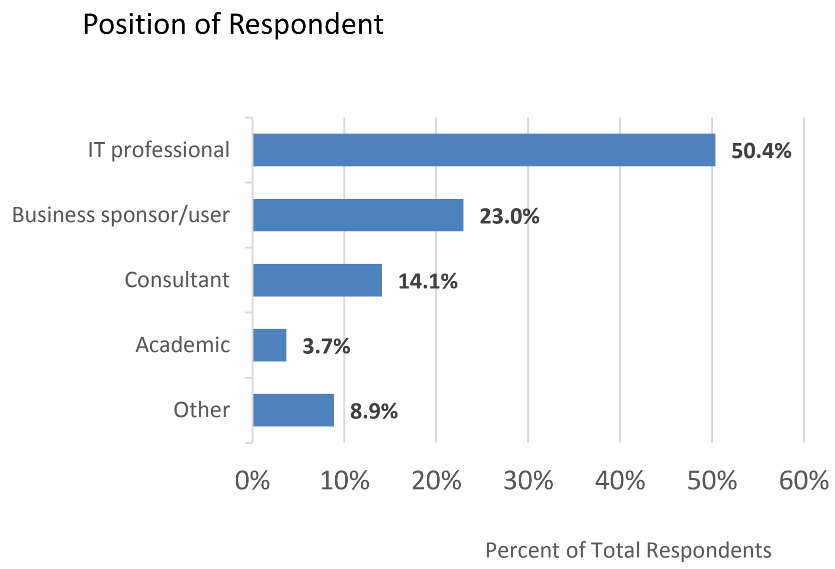


- The online distribution of the survey enabled the collection of data from around the globe. Nearly 70% of those completing the survey were located in the United States. India and Europe were the next most common sources at 9.6% and 8.9%, respectively.

- The research cohort was not required to answer this question; hence the 8.1% attributed to “Did Not Specify.”

- For analytic purposes the eight regions shown in the adjacent chart were consolidated to four:
  - Americas (69.6%): North and South America
  - EMEA (9.6%): Europe, the Middle East and Africa (not represented)
  - Asia/Pacific & India (12.6%): Asia, Australia and India
  - Did Not Specify (8.1%)

Demographics – Position



- The cohort pool represents a robust vertical cross-section of a typical organization. Respondents come from all levels — starting with the CEO and other members of the executive team — on both the technical and non-technical sides of the business. This brings together the perspective and experience of those who fund, develop, use and benefit from Hadoop-based solutions.

- The approximately 2:1 ratio of IT professionals to business sponsor/users highlights the keen interest that business managers have in Big Data and its possible applications. Furthermore, it underscores the need to have a substantive business question (regarding a problem or opportunity) as the basis of any Big Data initiative.

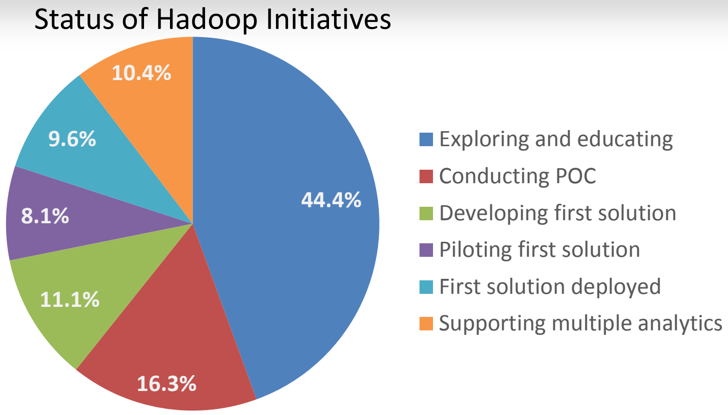
- Consultants (14.1%) bring added experience and insights from working with a multitude of companies across a range of industries.
  - Responses were received from top tier and global firms as well as regional and niche firms.
  - Consultants could be instrumental in the effective adoption and implementation of Hadoop-based solutions and the accelerated realization of value.

- “Other” encompasses those who did not provide role information.

Sand Hill Methodology for Qualifying the Progress

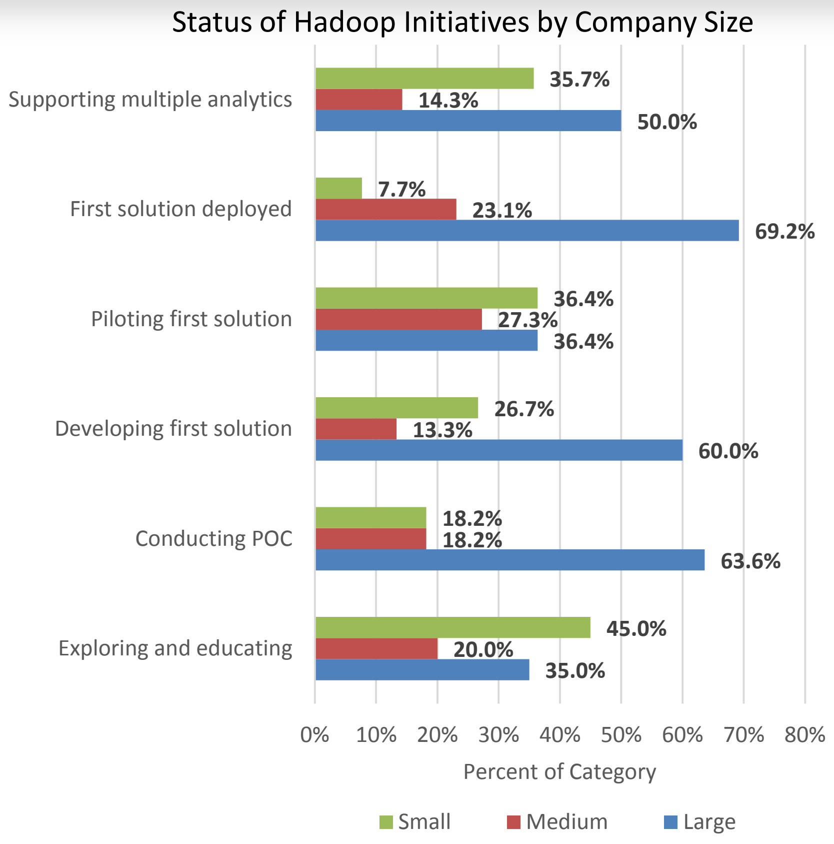
- Sand Hill Group’s Hadoop initiative qualification (HI-Q) process renders a three-dimensional view of the progress organizations are making toward realizing the full potential and value the platform can enable.
- The metrics defining the three dimensions are:
  - Hadoop-based initiatives status (from exploring and educating to supporting multiple business-critical analytics)
  - Uses of Hadoop (current and planned)
  - Data types used (from structure to unstructured, periodic refreshes to real-time feeds, etc.)

Steady Progress is Being Made, Though It's Still Early Days



- The first step in Sand Hill's HI-Q process was to gauge where companies are in the implementation and use of Hadoop-based solutions. The spectrum of possibilities spans:
  - Early Stage
    - Exploring and Education – Building competencies in Hadoop clusters and related technologies
    - Conducting a Proof of Concept (POC) – Confirming analytic effectiveness and clarifying analytic objectives
    - Developing the first solution
  - Intermediate Stage
    - Piloting the first solution with clearly defined commercial objectives and metrics for success
    - First solution deployed
  - Advanced Stage
    - Supporting multiple business-critical analytics
- As expected given the relative newness of the technology, 44% of respondents said they were in the process of understanding the technology and building the competencies required to build and operate business-critical solutions.
- Interestingly, 20% reported they have implemented one or more business-critical Hadoop-based solutions.
- The median phase for the overall population in this study is conducting a POC, which is relatively early in the process.

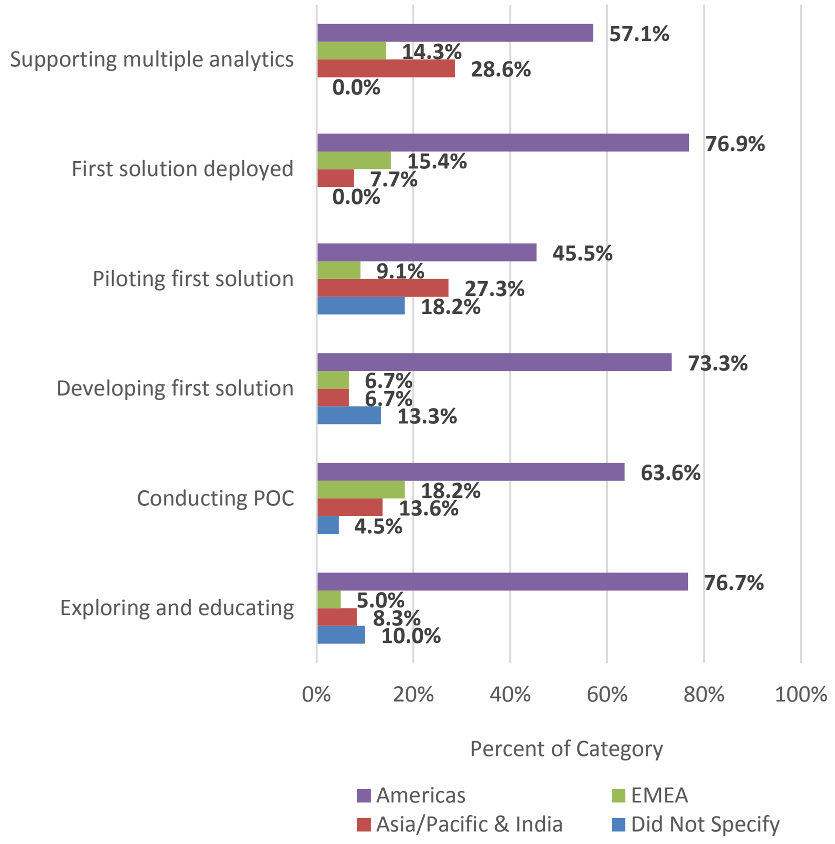
Steady Progress is Being Made, Though It’s Still Early Days



- Larger companies (1,000 or more employees) often have more available resources (people, time and budget) as well as vast quantities of data that need to be captured, managed, provisioned and interpreted. As a group on both a percentage and an absolute basis, large companies have advanced furthest of the three segments. Approximately 25% of these companies stated their Hadoop-based infrastructure currently supports at least one business-critical analytic application.
- Though small companies represent the largest share of the exploring and educating phase, driven by data-centric startup companies, they also have surpassed medium-size companies in advanced stages such as supporting multiple analytics and piloting the first solution.
- The median phase by company size is:
  - Small: Exploring and educating
  - Medium: Conducting a POC
  - Large: Conducting a POC

The Americas Lead with India Gaining Momentum

Status of Hadoop Initiatives by Region



- On a geographic basis the Americas, heavily weighted by the United States, rank number one in all phases of the adoption process.

- Asia/Pacific and India ranks second overall. This is powered primarily by India with its strong technology industry that provides advanced analytics and services to both national and international markets. In several instances, respondents from India reported they work for non-Indian companies, reflecting the continuing practice of leveraging the best and the brightest, irrespective of their physical location.
- The median phase by region is:
  - Americas: Exploring and educating
  - EMEA: Conducting a POC
  - Asia/Pacific & India: Developing the first solution
  - Did Not Specify: Exploring and educating

Mastering the Basics and Moving on to Advanced Applications

Most Commonly Reported #1 Uses of Hadoop
(Current vs. Future)

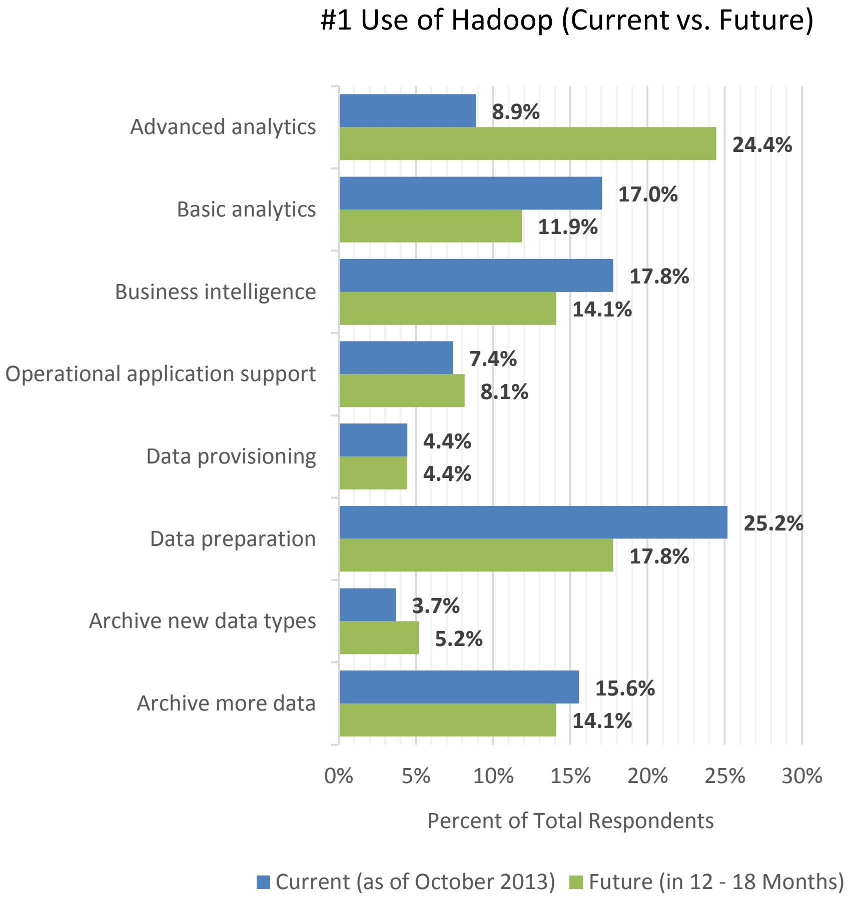
#1 Current Uses (as of October 2013)	#1 Future Uses (in 12-18 months)	Change From Current
Data preparation (25.2%)	Advanced analytics (24.4%)	—
Business intelligence (17.8%)	Data preparation (17.8%)	-7.4%
Basic analytics (17.0%)	Business intelligence (14.1%) Archive more data (14.1%)	-3.7% —

Top Uses of Hadoop (Current vs. Future)

Top Current Uses (as of October 2013)	Top Future Uses (in 12-18 months)	Change From Current
Basic aAnalytics (58.5%)	Advanced analytics (61.5%)	—
Business intelligence (48.1%)	Business intelligence (45.9%)	-2.2%
Data preparation (45.9%)	Data preparation (40.7%)	-5.2% —

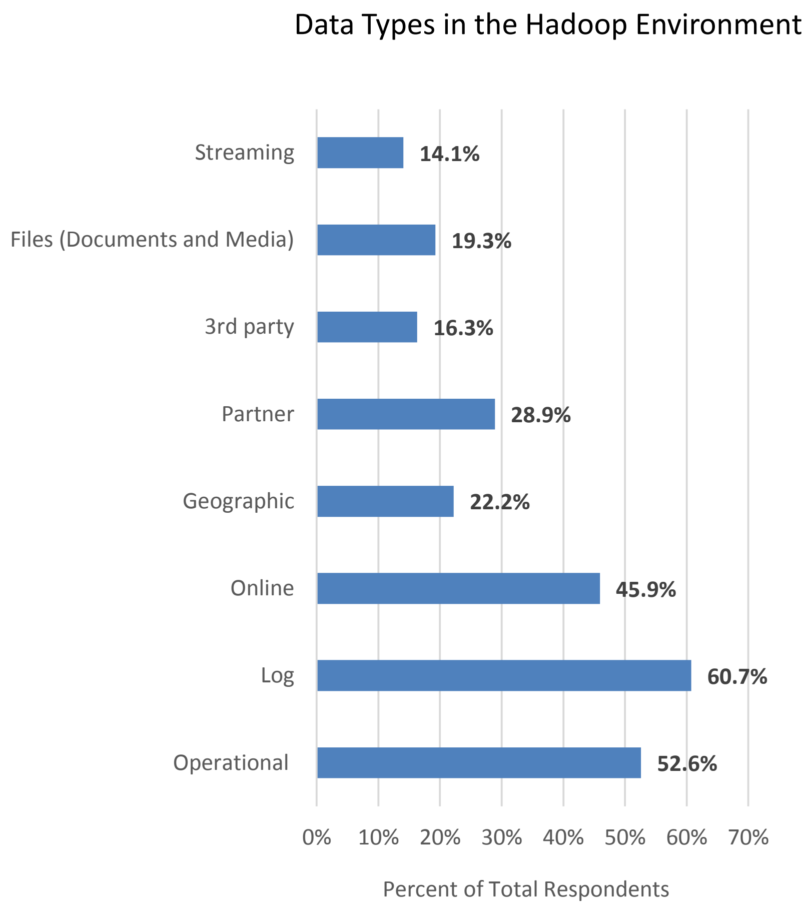
- The second step in Sand Hill's HI-Q process was to characterize the current and expected uses of Hadoop as a proxy measure for the level of technical sophistication and capabilities. The levels for this are:
  - Early capabilities:
    - Archiving more data
    - Archiving new types of data
    - Data transformation (ETL/ELT), data quality and preparation
  - Intermediate capabilities:
    - Data provisioning (delivering prepared data)
    - Business intelligence (e.g., reporting and dashboards)
    - Operational application support (running applications on Hadoop data)
    - Basic analytics (statistics, text, search, patterns, etc.)
  - Advanced capabilities:
    - Advanced analytics (e.g., risk, propensity, affinity, optimizations)
- To understand how organizations are evolving in their use of Hadoop, respondents were asked to identify, in order of importance, up to three current uses of Hadoop. Similarly they were asked to indicate their expected uses in 12-18 months' time.

Mastering the Basics and Moving on to Advanced Applications



- Consistent with the relative newness of the Hadoop platform, and reflecting that the bulk of the user community is in the exploring and educating phase, the current top uses of Hadoop tend to be:
  - Foundational
    - Data transformation (ETL/ELT), data quality and preparation
  - Supports or augments the organization's existing solution portfolio
    - Business intelligence (BI) – Providing a flexible, low-cost data-staging area for existing data warehouse/BI solutions. In addition, Hadoop can serve as a sandbox for further analyzing operational and other data.
    - Basic analytics – These application tend to use off-the-shelf visualization tools. In some cases, such tools are made available to end users for ad hoc analyses.
- Nearly one in four (24.4%) respondents stated that advanced analytics will be their #1 Hadoop-based initiative in the next 12-18 months. This almost-threefold increase over the 8.9% currently developing advanced analytics is emblematic of the larger shift towards initiatives that can have a transformational impact on the organization. It also conveys both the aggressive expectations for skill and experience development and the urgent need to mine the available data to improve business decisions and results.

The Data Does Indeed Tell the Story



- The third step in Sand Hill's HI-Q process was to evaluate the data used within the Hadoop environment. Type of data (real time or batch delivery), level of integration required and the complexity of the data provide another vantage point on the current analytic capabilities . The levels associated with data usage are:
  - Basic
    - Files (documents and media)
    - Logs from data servers, applications and other devices
    - Operational data from ERP, CRM and other enterprise systems
  - Intermediate
    - Online data from social, informational, interactive and user-generated sites
    - Data from business partners (such as supply chain, logistics and procurement)
    - Third-party data
  - Advanced
    - Streaming
    - Geographic
- The declining cost of storage and processing power incentivizes companies to save all data in the hope that it can be profitably mined for insights sometime in the future. Hadoop's ability to economically scale to handle massive volumes of structured and unstructured data further enables that.

The Data Does Indeed Tell the Story

- As with the systems interfacing with the Hadoop environment, the data hosted in the Hadoop environment confirms the applications in use today and reveals a glimpse of what companies may be planning to do in the near future.

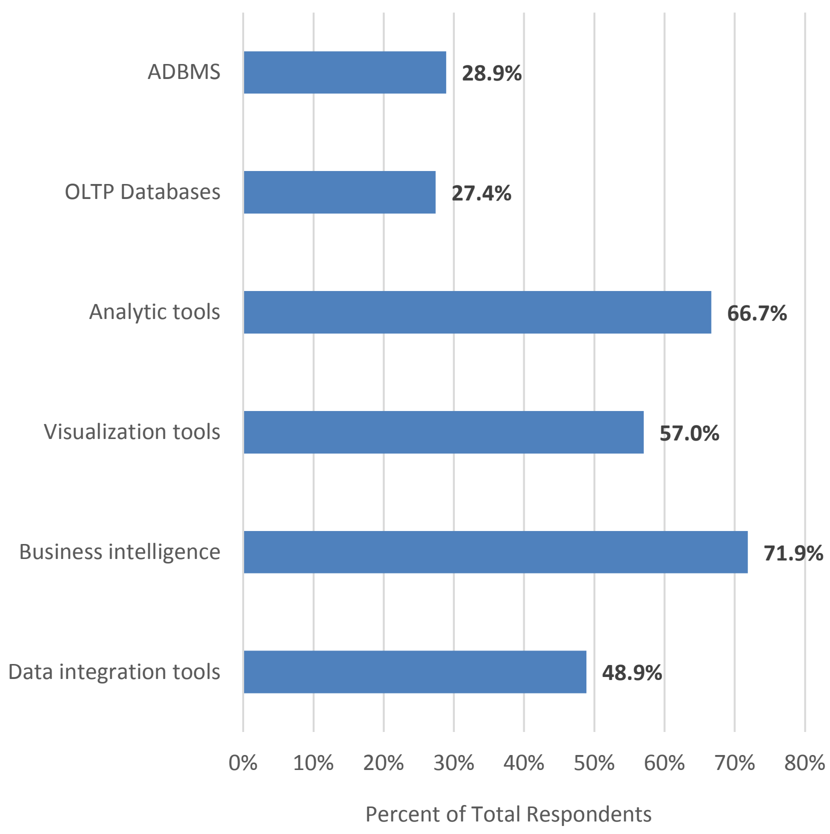
- Use of log data from servers, applications and other devices (60.7%) is the most widely stored data type in the Hadoop environment. Such data can be used with custom or third-party analytics to, among other things, more effectively operate the IT infrastructure, manage costs and verify SLAs.

- Operational data from ERP, CRM and other enterprise systems (52.6%) can be used on its own or combined with other data available to the organization — including online data from social, informational, interactive and user-generated sites (45%) — in a Hadoop-based analytic sandbox, supporting ad hoc analyses not easily conducted in a traditional BI context.

- In terms of future applications enabled by data collected today, geographic data from geo-specific applications and/or mobile devices and apps (22.2%) and streaming data (14.1%) suggest an accelerating move from the back office, offline analytics to real time and interactive solutions. Doing so will require new enterprise-class features such as enhanced online and high-availability features to be embedded in the Hadoop ecosystem.

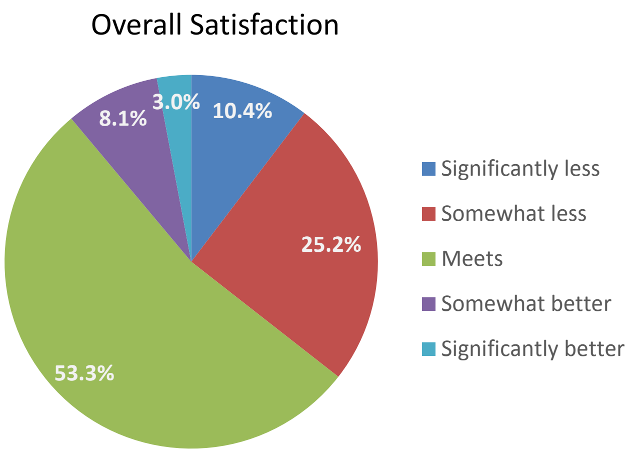
Data Access Priorities Correlate to Today's Top Uses of Hadoop

Other Systems Used in Conjunction with Hadoop



- Since data needs to be transferred into and out of the Hadoop environment, it often must interface with a variety of other systems. The top three most-referenced systems — business intelligence (71.9%), analytic tools (66.7%) and visualization tools (57.0%) — map directly to the current top uses of Hadoop: data transformation (ETL/ELT), data quality and preparation (25.2%); business intelligence (17.8%) such as dashboards, reports and visualizations; and basic analytics (17.0%) such as statistics, text, search and patterns.

Satisfaction is not a Given

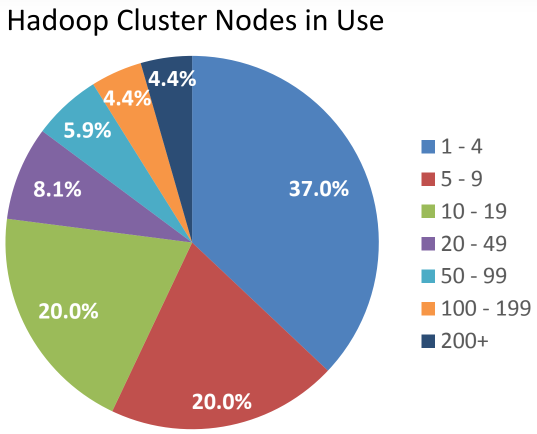


- On the positive side, 53.3% of the cohort reported that their Hadoop-related experience solidly met expectations. This suggests that the Hadoop platform is well on its way to emerging from the hype and its associated disappointment.

- On the less-than-positive side, there is a significant gap between those who believe the results of their Hadoop/Big Data initiatives were better/less than expected:
  - 35.6% of companies indicated results were somewhat or significantly less than expected
  - Only 11.1% of respondents said results achieved were either somewhat or significantly better than expected

- In the near term, creatively addressing the skills and experience shortages will create improved outcomes and greater satisfaction.

Scale Your Hadoop Cluster as You Go

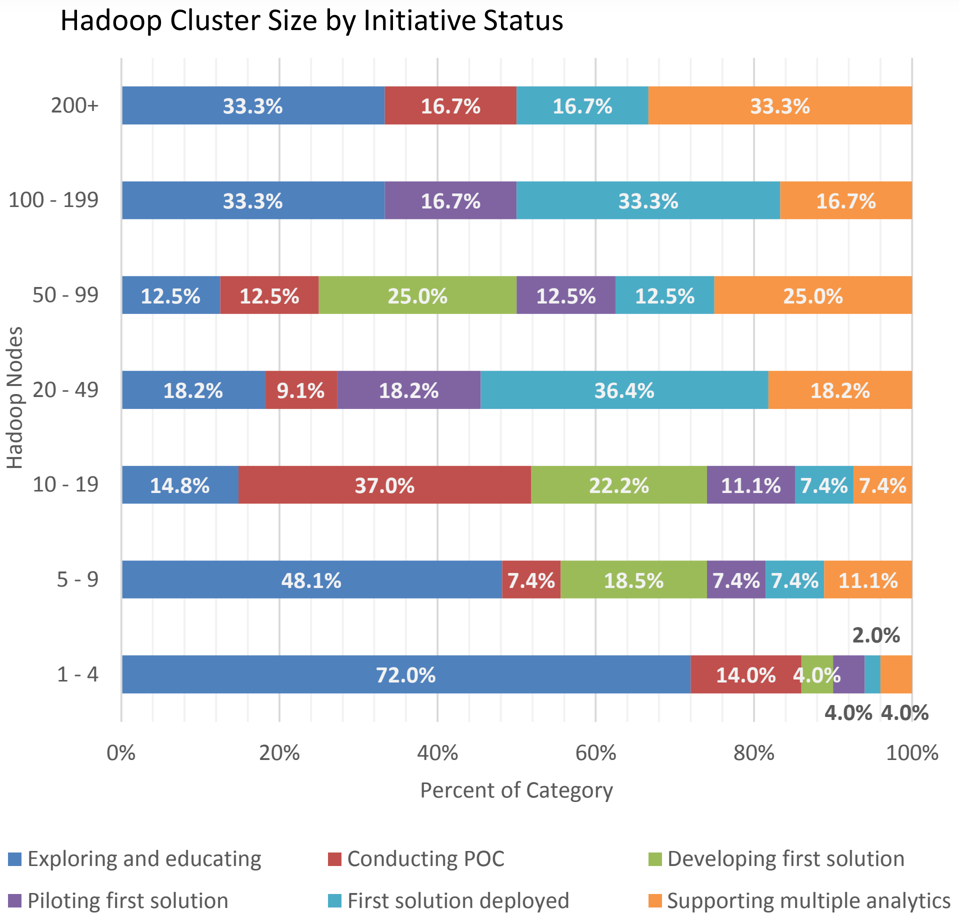


- Scalability of Hadoop allows the rapid processing of massive volumes of data. The tools in the Hadoop platform reduce the complexity associated with managing clusters of any size.

- The median size of Hadoop clusters currently in use is in the five- to nine-node range. This is indicative of the fact that approximately 60% of those surveyed are at an early stage of the adoption process (exploring and educating, and conducting a POC).

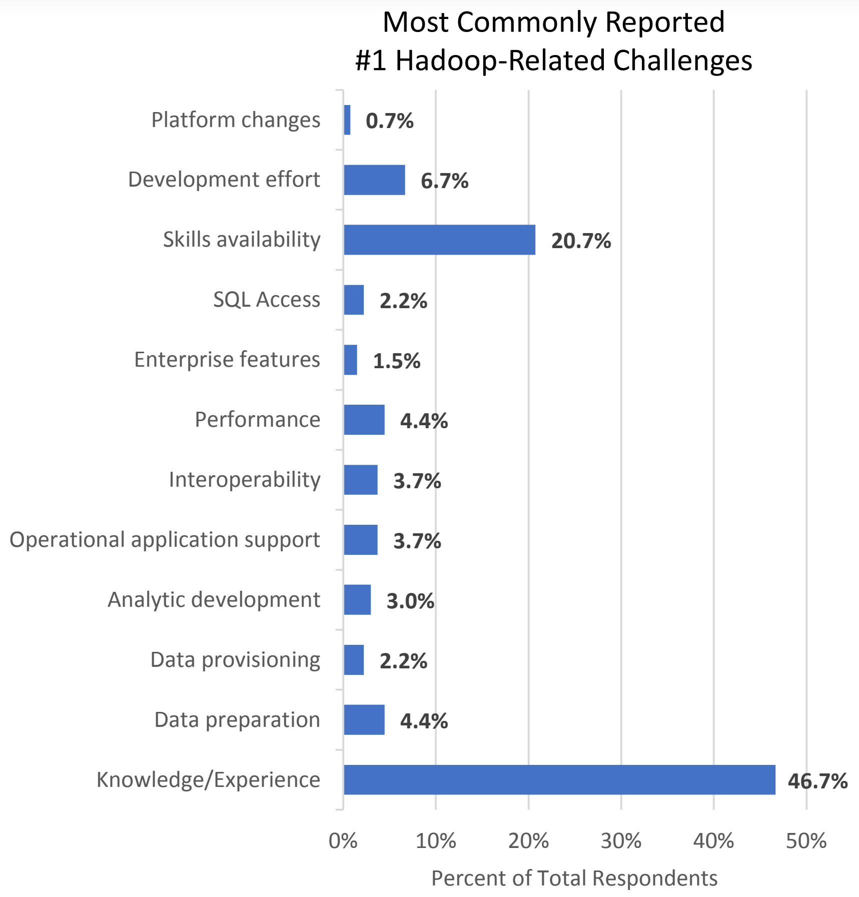
- For the purposes of this study, Sand Hill Group does not distinguish between Hadoop clusters located within a company's data centers or accessed via the cloud. Cloud-based Hadoop environments offer companies an economical way to establish experience with Hadoop clusters of all sizes without the typical up-front investments.

Scale Your Hadoop Cluster as You Go



- Cluster size tends to be aligned closely with the purpose of each stage in the process of developing and deploying Hadoop-based solutions.
- Smaller cluster sizes tend to be used extensively as an educational platform for experimenting and developing the necessary competencies. A mere 6.0% of the companies with clusters of four or fewer nodes report using the system for production applications.
- Conversely, it is common that 50% of cluster configurations containing 20 or more nodes are used for production applications. These systems are used on a much more limited basis for early-stage activities such as exploring and educating.

Second-Tier Challenges Related to Hadoop Must Remain on the Radar Screen as Today's Top Issues are Resolved



- To identify the negative drivers behind the satisfaction ratings, respondents were asked to specify, in order of significance, up to three challenges they encountered in the course of using Hadoop.
- The three most commonly reported #1-ranked challenges are the current level of knowledge and experience with the platform (46.7%), the availability of the required skills (20.7%) and the amount of technology development and engineering required (6.7%).
- Though the development effort required ranks a distant third today, it foreshadows a significant barrier to the adoption of the platform for more advanced applications. In fact, the next tier of issues — data preparation, performance, operational application support and interoperability — may prove to be larger-than-expected issues down the road if left unaddressed.
- At this time, there is a well-known scarcity of experienced data scientists, solution architects, analytic developers and other technology experts required to implement and operate Hadoop-based applications. Given the demand for these individuals, they are often recruited by brand-name companies (e.g., Google and Facebook) or startups where they can be part of leading-edge initiatives, have access to a plethora of employee benefits and receive a compelling compensation package.

Second-Tier Challenges Related to Hadoop Must Remain on the Radar Screen as Today's Top Issues are Resolved

- Adding to the talent-pool frustration, it is often a time-consuming process to cultivate the capabilities internally. Frequently redirecting existing IT staff proves to be a skills and cultural mismatch. As an example, it may be challenging for those steeped in "small data" analytics and business intelligence to embrace the Big Data mindset and methodologies.
- The ability to maximize the productivity and effectiveness of limited development resources is critical. Visualization, analytic and other types development tools are available today from a number of vendors; however, they are not as robust in terms of features and stability as those available for business intelligence, Web and traditional software development. As mentioned previously, this is expected to be a growing pain point until the available tool set expands and matures further.

Bradley Graham

Bradley Graham is the executive director of Carpe Datum Rx, a thought leadership forum for the business application of advanced analytics. He is also the vice president of marketing for Amberoon Inc., a provider of data-driven business perspective solutions. Prior to that Bradley was the head of product management for Fair Isaac Corporation's global scoring and analytics business (which includes the FICO Score) and was on the executive leadership team of several technology companies in the broadband, digital TV, document management and Internet services markets. He can be reached at bgraham@amberoon.com or @CarpeDatumRx.

M. R. Rangaswami

M.R. Rangaswami thrives on creating influential communities that achieve global impact. Using his skills as a critic, cheerleader and facilitator, M.R. has united leaders and advanced strategic initiatives in the fields of software, corporate sustainability and entrepreneurship. He is the publisher of SandHill.com, the premier online destination for strategic information on the Big Data, cloud and mobile software ecosystem. Contact M.R. at mr@sandhill.com.

Partners

Carpe Datum Rx is a thought leadership forum for the business application of advanced analytics technologies.

www.CarpeDatumRx.com

The Hive works with entrepreneurs to create companies that use data to innovate products and gain advantage in a competitive market.

www.HiveData.com