Wikibon Big Data Analytics Adoption Survey, 2014-2015

Hadoop Takes Aim at the Enterprise

Analysts:

Jeff Kelly

David Floyer

Ralph Finos

Wikibon

Contents

LIST OF FIGURES

Executive Summary

Big Data analytics is a popular subject of debate and frequent topic in the media. Hardly a day goes by without an article or two about the use of Big Data analytics in one large enterprise or another. Judging by media reports, Big Data analytics is as easy as collecting some data, deploying an application or two and waiting for the "game-changing" insights to start rolling in.

The reality is Big Data analytics is hard. For every success story highlighted in The Wall Street Journal or trade publication, there are many others enterprises struggling to get value out of their investments in Big Data analytics technologies, services and staff. We know this anecdotally, but there is also hard data to back up this claim.

A September 2013 survey of the Wikibon community[1] found that 48% of Big Data practitioners have yet to realize the full value of their Big Data analytics investments. On average, respondents reported realizing just $.55 in return for every dollar invested. Still, expectations are high. Respondents said they expect to ultimately realize a return of $3.50 per dollar invested in Big Data Analytics over the next three to five years. Clearly, these practitioners and organizations have a long way to go.

So what, exactly, are the main barriers to successful Big Data Analytics projects? That was one of the main questions Wikibon set out to answer with the Wikibon Big Data Analytics Survey, 2014. But to truly understand the barriers and challenges facing Big Data practitioners, we also need to understand the context in which they occur. This means understanding the state of Big Data analytics projects across vertical industries. This includes identifying adoption rate by industry, the technologies in use, the types of data being analyzed, and where in the lifecycle projects stand.

It is also important to understand the challenges facing Big Data analytics practitioners based on role. Our premise is that the challenges facing infrastructure professionals relative to Big Data analytics, who are tasked with standing up and maintaining the technology, are very different from the obstacles facing applications developers, for example.

With these objectives in mind, Wikibon launched a web-based survey of Big Data practitioners in May of 2014. We asked survey respondents to describe the state of their existing Big Data Analytics projects, the use cases involved, the technologies/services/data involved, their personal roles in projects, and the main barriers and challenges they and their colleagues encountered along the way. What follows is the analysis of the survey data. Our hope is that this analysis provides both a big picture view of the state of Big Data analytics in the enterprise in 2014 as well valuable insights for Big Data analytics practitioners and vendors regarding common barriers to success.

1 Kelly, Jeff. "Enterprises Struggling to Derive Maximum Value from Big Data." 19 Sep. 2013

Methodology and Demographics

Understanding the methodology of the survey is critical to interpreting the resulting analysis. Our methodology, as well as survey respondent demographic information, is as follows.

Wikibon conducted a web-based survey of 303 Big Data analytics practitioners in the US in May 2014. Survey respondents were asked to characterize their understanding of Big Data analytics at the outset of the survey. Those who responded they were "somewhat familiar" or "very familiar" with Big Data analytics, as defined below, were asked to proceed with the survey. Those that said they were "unfamiliar" with Big Data analytics were terminated from the survey.

For the purpose of this study, we defined Big Data analytics projects as those that:

- Leverage non-traditional data management tools and technologies such as Hadoop, NoSQL, or MPP analytic databases, and/or …
- Involve the analysis of multi-structured and/or unstructured data such as clickstream, text, log file, and social media data.
- Big Data projects, for the purpose of this survey, do not include projects solely involving the use of relational databases or otherwise "traditional data management technologies" to collect, process, store and analyze structured data associated with legacy systems such as CRM and ERP applications.

The survey further asked respondents to identify which industry they worked in, their role in the enterprise generally and role specific to Big Data analytics projects, employee count and annual revenue of their enterprise.

As a result, Wikibon obtained a broad distribution of enterprise types, led by IT technology providers, healthcare, manufacturing, banking & finance, and retail **[Figure 1**]. The enterprises represented ranged from companies with at least $10 million in annual revenue up to $1 billion plus. The median company size was between $100 million and $500 million in annual revenue with between 1,000 and 5,000 employees.

Respondents' level of responsibility ranged from managers to c-level executive. Respondents were also asked to identify their roles **[Figure 2]** related to Big Data analytics projects by selecting one of the following personas:

- Business user (i.e. A line-of-business professional who uses dashboards and other visualizations to understand Big Data).
- Business analyst (i.e. A departmental power-user who conducts analysis of various Big Data sets with tools such as Excel and SPSS.)
- Application developer (i.e. A developer who builds applications that leverage Big Data analytics such as predictive models and algorithms.)

- Data scientist (i.e. An advanced analytics professional who conducts sophisticated analytics and develops predictive models/algorithms on large volumes of "messy" Big Data.)
- Infrastructure administrator (i.e. A datacenter professional who manages infrastructure and hardware associated with Hadoop, NoSQL database and other technologies that support Big Data analytics projects.)

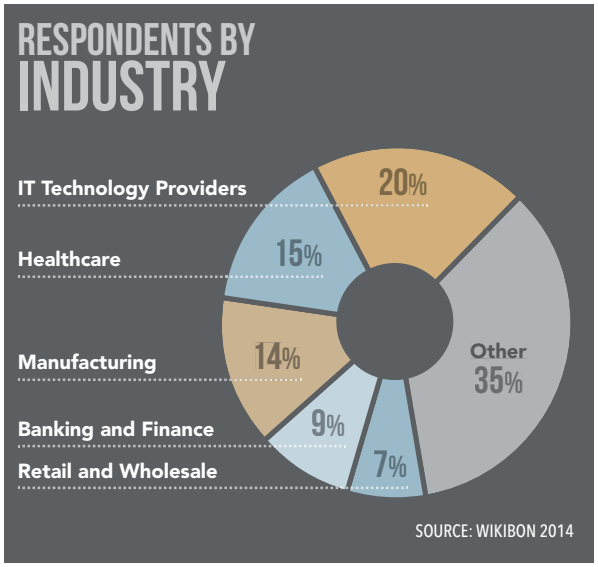

FIGURE 1

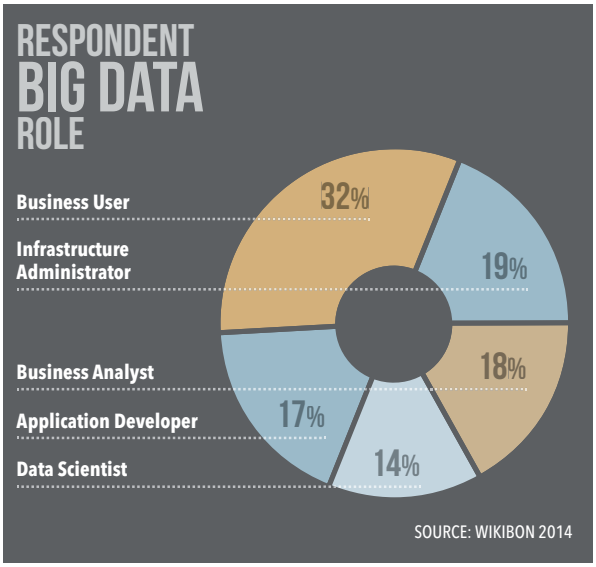

FIGURE 2

ANALYSIS: Based on the respondent profile and their understanding of Big Data analytics, it is clear that the resulting analysis represents the state of Big Data analytics among relatively early adopters. This is an inevitable result of studying this topic. With Big Data analytics technologies and approaches still relatively immature, those enterprises and practitioners that are evaluating or have deployed Big Data analytics projects are by definition early adopters. This is an important piece of information to keep in mind when considering the results of the survey.

TOP TEN FINDINGS

1. Vast majority of respondents believe Big Data Analytics is critical to the success of their respective enterprises.
2. Big Data Analytics is a transformational movement, not just a tactical tool or set of technologies.
3. New approaches (Hadoop, NoSQL, open source and cloud) are already disrupting traditional data management an analytics approaches in the enterprise.
4. Most common data volume in Big Data Analytic deployments range between 50 &100TB and most deployments involve the blending of two or more data sources.
5. Big Data Analytics use cases and types of data involved are wide and varied.
6. Most deployments are still in evaluation and proof-of-concept phases, but a tipping point to production is imminent.
7. Enterprises are looking for guidance with professional services engagements a given for the majority of Big Data Analytics projects.
8. Data integration, data transformation and integrating with existing infrastructure are the biggest technology barriers to success.
9. Selling the "value" of Big Data Analytics to the business, getting stakeholders to agree and uncertainty regarding compliance & privacy are the biggest non-technology barriers to success.
10. IT practitioners have a much rosier perception of the relative success of Big Data Analytics projects versus non-IT practitioners indicating a misalignment between the two.

State of Big Data Analytics in the Enterprise

In order to fully understand the barriers to success facing Big Data analytics practitioners, we must understand the state of deployments in the enterprise. Following is an assessment of the state of Big Data analytics in the enterprise based on the survey data.

A vast majority of our survey respondents believe Big Data analytics will have a significant impact on the future competitiveness of their organizations. 46% believe that "Big Data analytics is the new source of competitive advantage," while another 46% believe that "Big Data analytics is/will be an important compliment to our existing data warehouse and business intelligence practice." That compares to just 8% of respondents who believe Big Data analytics is simply a "nice to have" set of technologies/capabilities, and just 1% that feel Big Data is little more than a buzzword **[Figure 3]**.

Clearly, Big Data analytics is viewed as a transformational approach to harnessing the value of data, not just a set of tactical tools or technologies.

CASE STUDY:

TrueCar is an online service that lets consumers and dealers search for and compare prices on both used and new cars. "Our mission is to bring transparency to complex marketplaces," says John Williams, Senior Vice President of Platform Operations at TrueCar. "Buying a car is certainly a very complicated transaction and it's something involving a lot of data." The data people typically consider when evaluating cars include car makes, models, years, colors, options and pricing data. For used cars, the history of the car, such as accident information, is also very important. As are images. "Obviously, its very important that you want to buy something you can see." TrueCar processes over one million car images per day.

TrueCar uses Hadoop, specifically the Hortonworks Data Platform, to manage and process all that data – over 2.5 petabytes. Williams says his company simply couldn't do what it does without a framework and technology approach such as open source Hadoop. From a cost perspective, "we realized the economics of storage has changed. Storage is now so inexpensive. You can effectively now store all your data forever. And when you really think that through, you know have an infinite timeline where you can go back and remonetize the same piece of data many, many times in the future."

While Big Data analytics allows practitioners to improve existing applications, it's the new capabilities, which approaches like Hadoop enable, that makes Big Data so impactful, says Williams. "It's not about doing traditional things bigger and faster. Big Data is really transforming our business on a fundamental level."

While most our survey respondents believe Big Data analytics holds significant promise and will play an important role in the future success of their respective enterprises, it is still early days. The majority of respondents, 69%, reported that their organizations were either in the evaluation stage of Big Data analytics projects – meaning they were still evaluating use cases, technologies and vendors – or

projects were in the pilot/proof-of-concept phase. Less than a third, 31%, reported they had at least one Big Data analytics project in production supporting mission-critical applications **[Figure 4]**.

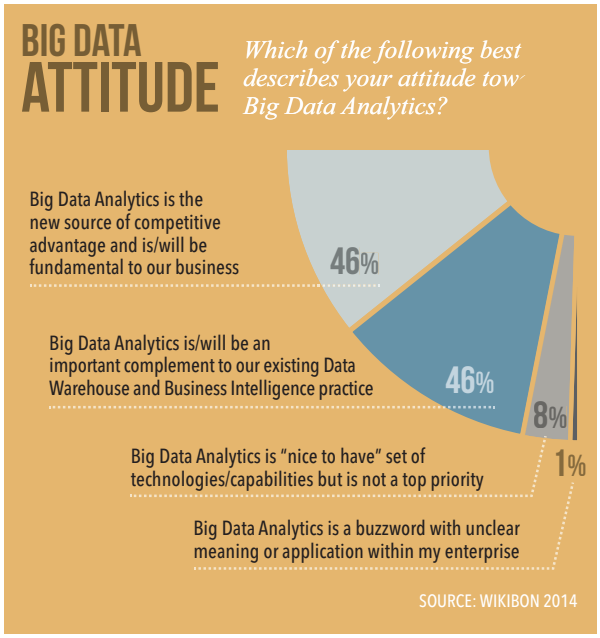

FIGURE 3

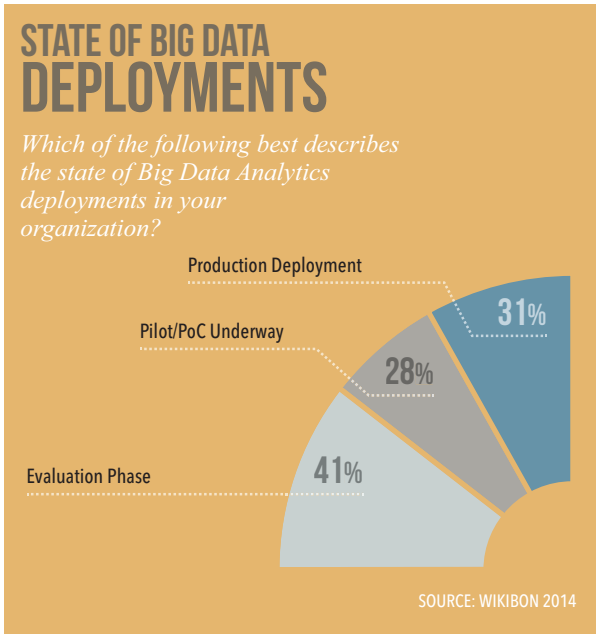

FIGURE 4

Use cases reported were wide and varied. Topping the list of Big Data analytics use cases was competitive analysis, followed by IT equipment operations support, data transformations, risk management and workflow optimization.

ANALYSIS: The main takeaway from the varied use cases, combined with the varied vertical industries represented by survey respondents, is that Big Data analytics is applicable across use cases and vertical industries. As Wikibon wrote back in 2011 in its Big Data Manifesto [2], there is no industry that won't be impacted by Big Data analytics, and these survey findings validate that contention.

Most respondents reported positive results, at least initially, related to Big Data analytics projects **[Figure 5]**. 41% reported they have realized the full value of their investment in Big Data technology, services and staff, with 58% reporting just partial return on Big Data investment but moving in the right direction. However, there is a distinct schism between IT and the business when it comes to characterizing the relative success or failure of Big Data analytics projects. Of the survey's respondents that work in IT departments, 54% reported that their enterprise or organization had realized the full value of its Big Data analytics investment. Conversely, just 18% of non-IT respondents felt that full value had been realized.

2 Kelly, Jeff. "Big Data: Hadoop, Business Analytics and Beyond." 29 Sep. 2011

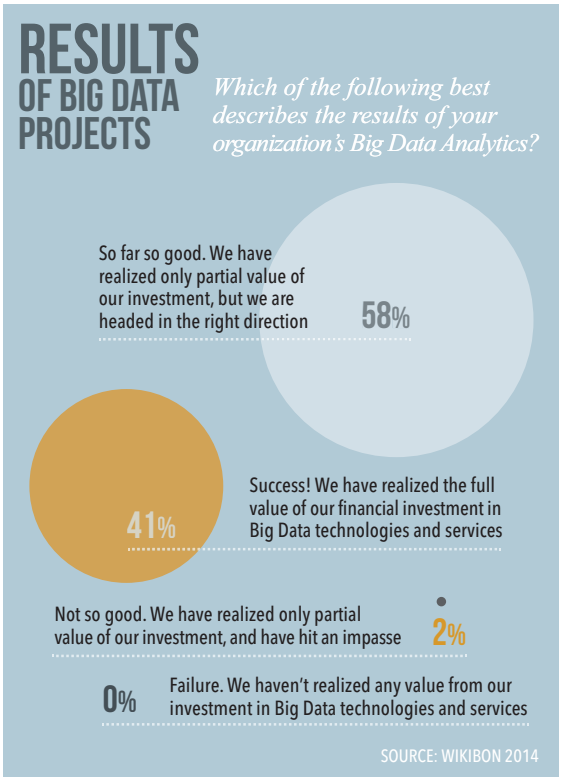

FIGURE 5

ANALYSIS: This high proportion of positive results overall (those reporting they had achieved full value of Big Data analytics investments or partial return but moving in the right direction) is likely due in part to the relatively low-level of investment made to-date in Big Data analytics technologies, services and staff. It is none-the-less encouraging that most practitioners feel positively about such investments.

But the divide between IT and non-IT respondents is striking. Wikibon believes this schism is due in part to the criteria used to judge success by the two groups. IT department practitioners are tasked, in part, with standing up and maintaining the technologies and infrastructure used in Big Data analytics projects. Non-IT practitioners are more concerned with the analysis of Big Data and gaining actionable insights. The survey results indicate that IT department practitioners feel they are doing their jobs well, but the non-IT practitioners (i.e. line-of-business) are struggling to extract insights from Big Data. This is, to a degree, to be expected considering the lifecycle of deployments. The technology must first be deployed and operating successfully before data can be analyzed for insights. This is yet another indication of just how early we are in the days of Big Data analytics.

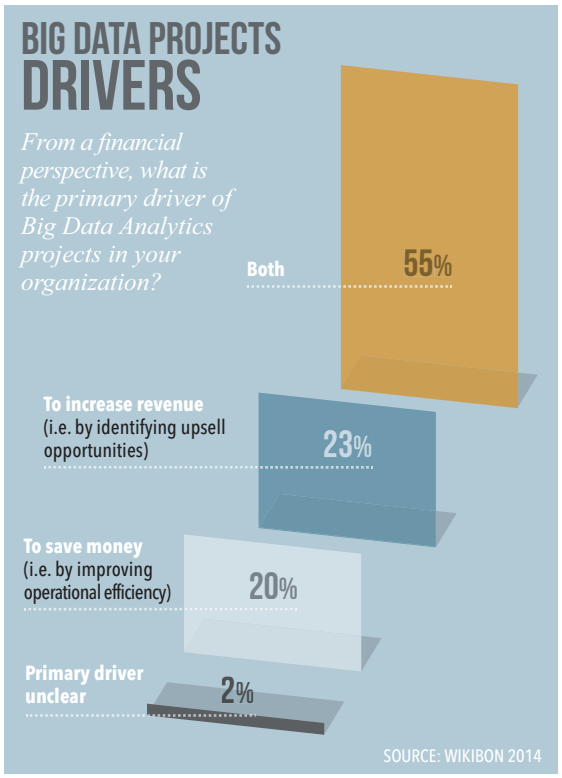

FIGURE 6

As for financial drivers of Big Data analytics projects, 20% of respondents reported that saving money was the primary driver, with 23% saying increasing revenue was the primary driver. A majority, 55%, reported both saving money and increasing revenue were primary drivers of their organizations' Big Data Analytics projects **[Figure 6]**.

ANALYSIS: The 20% of respondents that said saving money is the primary driver are at risk of overlooking the revenue generating potential of Big Data Analytics which manifest themselves in a number of ways – identifying cross-sell and upsell opportunities, better advertising and marketing campaign targeting, developing net new products and lines of business, etc.

As part of the survey, we asked a series of questions to get a better understanding of the types and volume of data being analyzed and the tools and technologies in use. Regarding types of data involved in Big Data analytics projects, social media data was the most common data involved. Structured

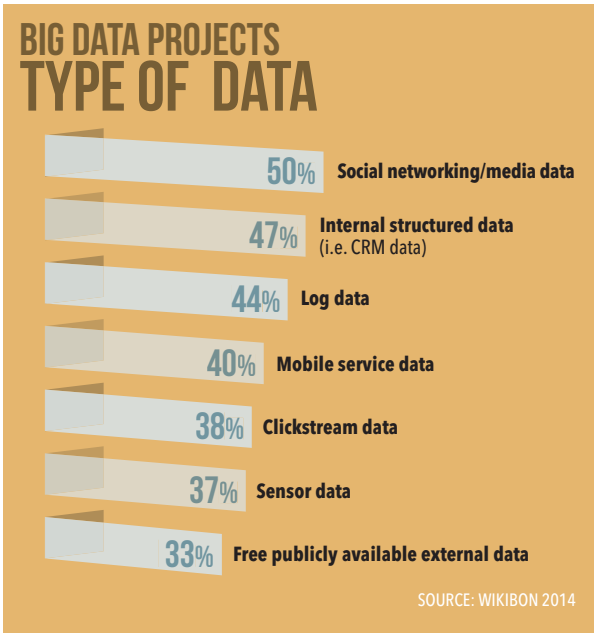

FIGURE 7

data from internal systems was the second most common type reported by respondents. Log data, mobile device data and click-stream data rounded out the top five types of data involved in Big Data analytics projects **[Figure 7]**.

In a related question, we asked respondents if they had integrated two or more types of disparate data together for analysis. Merging disparate data sets can provide new and unexpected analytic insights and represents the real promise of Big Data analytics. Encouragingly, over 60% of respondents reported they had indeed integrated two or more disparate data sets, with 32% planning to do so. However, as noted later in this report, data integration is also one of the top Big Data analytics-related challenges, according to respondents.

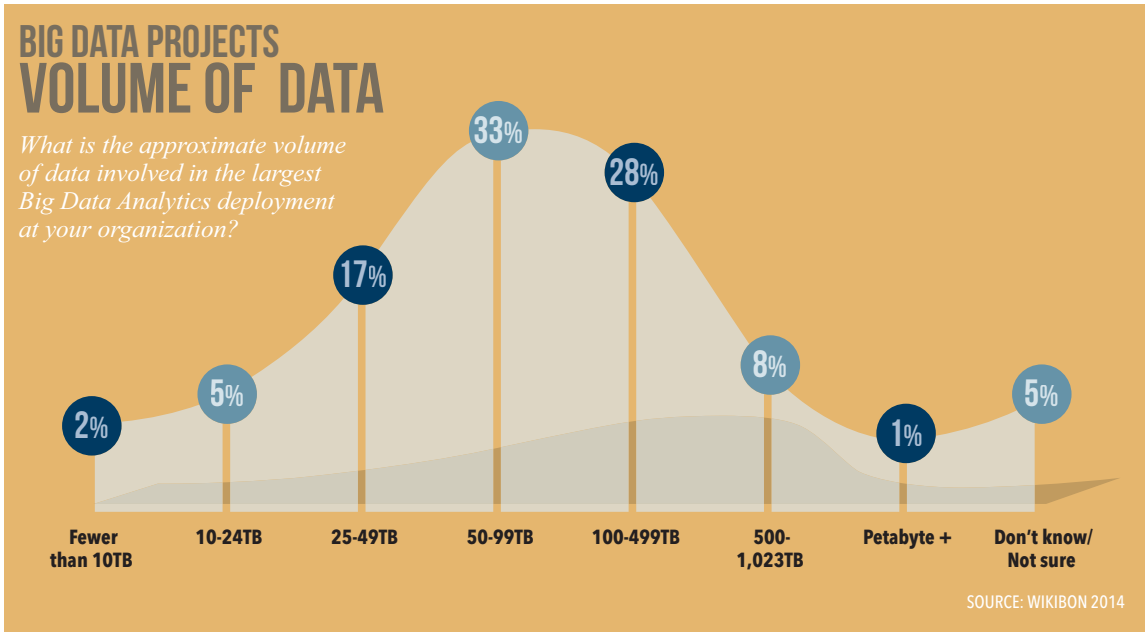

FIGURE 8

As for data volume, the survey indicates the average Big Data analytics deployment is 265 terabytes in size. The most common deployment was between 50 and 99 terabytes of data (33%) **[Figure 8]**.

ANALYSIS: Practitioners are likely exploring social media data such as Tweets and Facebook posts to better understand customer sentiment and to identify customer service-related issues. While most think of unstructured data when thinking of Big Data, it is logical that many practitioners begin new

projects looking at data they already have under management in relational systems. When it comes to data volume associated with Big Data analytics projects, Big Data really is big compared with transactional systems. Still, while Big Data is often equated with petabytes or more of data, just 1% of respondents reported deployments of petabyte scale.

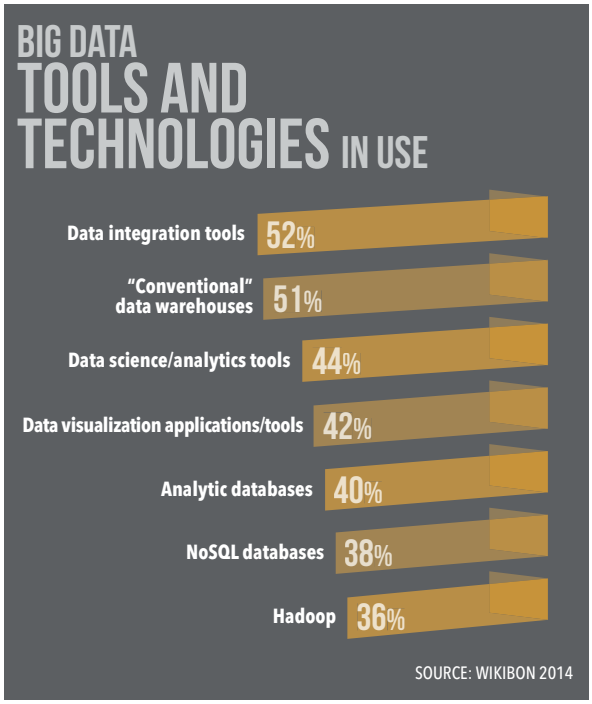

FIGURE 9

Which technologies and tools qualify as Big Data is a hot topic of debate. What is clear from the survey is that "traditional" technologies and tools are being applied to Big Data analytics workloads in conjunction with more emerging technologies such as Hadoop and NoSQL databases. While Hadoop and NoSQL are important technologies in Big Data analytics deployments, traditional data warehouses continue to play a role, at least for the time being. Not surprisingly, data integration technologies are also critical for moving data between systems when required and to assist in data transformations **[Figure 9]**.

The concepts of cloud computing and Big Data are often mentioned in the same breathe. Indeed, considering the complexity of the technology and the potential scale of infrastructure required to support Big Data analytics projects, the public cloud must be a consideration for many enterprises. However, as noted by many in the media and elsewhere, there are a numbers of barriers to wide-scale deployment of Big Data analytics in public cloud environments, not least of which being security and privacy concerns related to sensitive data sets.

CASE STUDY:

A large cable and Internet service provider is using Hadoop to manage and analyze large streams of customer usage data coming from set-top boxes and mobile devices. While Hadoop enables the company to store, process and analyze the data in a cost and time-effective way, Hadoop isn't the only technology in use to support this use case. In addition, the company is using a variety of data integration tools – both Hadoop-related tools and more traditional data integration software – to get data into Hadoop and to transform it into suitable form for analysis once there. Said one Big Data analytics practitioner at the company: "Data integration has to happen some way. At some point you have to pay that cost. Unless your data is landing in HDFS in a beautiful clean format that you can then leverage and run Hive on it straightaway without having to do anything else on it, then you're extraordinarily lucky and, yeah, that will work. Otherwise you're going to have to do some [data]

Wikibon asked respondents if they used the public cloud for any part of Big Data analytics deployments, both pilot projects and production deployments. To our surprise, 58% of respondents reported using the public cloud for Big Data analytics projects **[Figure 10]**. Of those, over half reported utilizing public cloud services to support production deployments. Another 26% of respondents reported that they were not currently using the public cloud for Big Data analytics projects but planned to do so in the future.

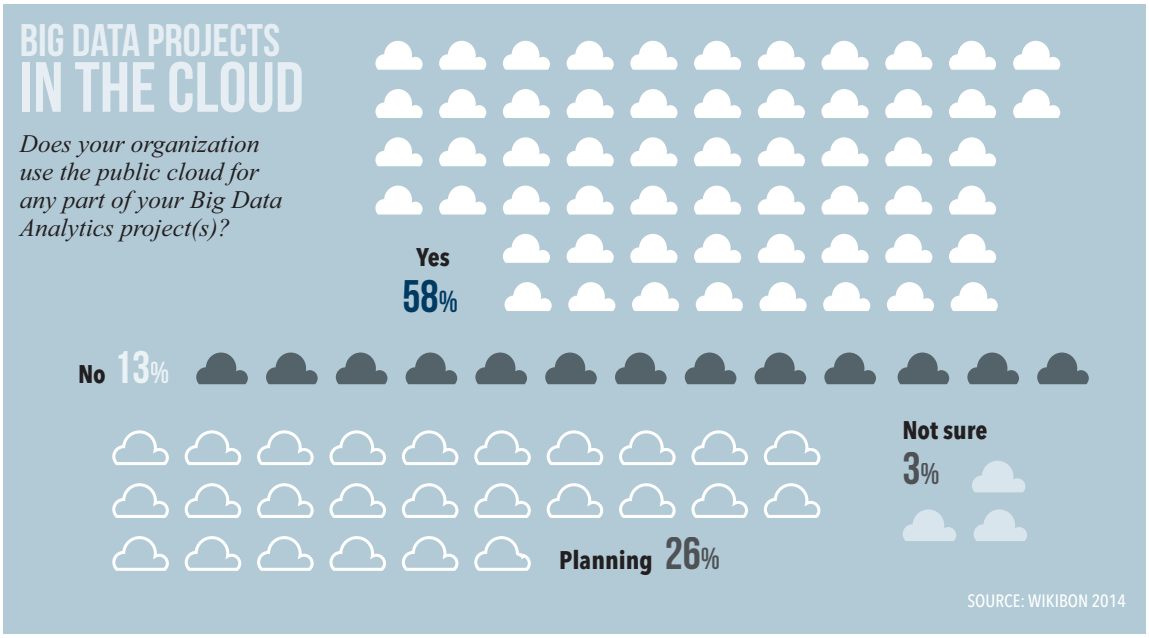

FIGURE 10

ANALYSIS: Maintaining the infrastructure to support Big Data analytics projects is a complex and costly affair. Even early adopters, who tend to have higher levels of internal expertise than mainstream adopters and laggards, recognize they can't do it all themselves. Offloading parts of Big Data analytics projects to cloud services allows enterprises to focus more on the value offered by Big Data analytics – namely gaining actionable insights to move the business forward – and less on maintaining infrastructure.

CASE STUDY:

Interactions Marketing is in the "retail food sampling and demonstration events" business. The company works with customers in the retail and CPG industries to put on events as small as one-on-one in-store food sampling stations to large outdoor interactive campaigns. The company relies heavily on Big Data analytics to support its operations, according Abhi Beniwal, Interaction Marketing's Senior Vice President of Global Information Technology. Beniwal says the company integrates data from its retail and CPG customers – including store location data, historical sales data and loyalty card-related data – with consumer data from third-party providers such as Nielson and publicly available weather and demographic data. It performs analytics on the merged data to help it make decisions about what types of campaigns to run based on customer objectives – namely to attract customers and drive sales. And it does all of this in the public cloud.

Interactions Marketing utilizes Google Big Query to process and analyze the data, along with Tableau Online to visualize it. Beniwal says the company considered performing Big Data analytics inside its own data center using Hadoop, but it wasn't feasible. Despite that data is such a critical part of its business, Beniwal says the learning curve associated with Hadoop for his staff was prohibitive. "We looked at [Hadoop] but we just couldn't figure out how we could make it work with the resources we had, the skill set we had, and the time frame we were looking at." A cloud service such as Big Query removed these burdens, Beniwal says, and allows the company's Data Scientists to focus on analyzing the data for insights instead of maintaining complex and growing infrastructure. And as the company looks to add new data sources from outside its firewall to its Big Data analytics practice – the company and its retail and CPG customers are beginning to experiment with sensor data, for example – the public cloud makes integrating that data simpler.

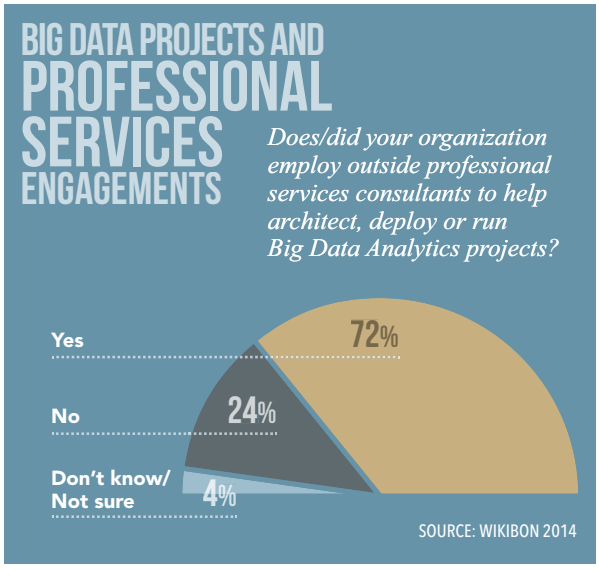

FIGURE 11

Wikibon asked survey respondents if their respective enterprises engage professional services firms and systems integrators to support Big Data analytics projects. A vast majority of respondents – 72% - reported employing outside consultants to support Big Data analytics projects **[Figure 11]**.

ANALYSIS: This is not surprising considering the relative immaturity and complexity of related Big Data analytics technology. Like cloud services, professional services and system integrators offer practitioners much needed support when deploying and maintaining complex infrastructure. But they also provide helpful guidance related to selecting the most valuable Big Data analytics use cases and transforming corporate culture to embrace data-driven decision-making.

Barriers to Big Data Analytics Success

Wikibon asked respondents to identify the top challenges and barriers to success they face relative to Big Data analytics projects. Respondents were given over two-dozen choices to select from, including technology and non-technology-related barriers to success, and were offered the option to write in their own selection.

Below are the findings for all survey respondents.

The top five technology barriers to successfully realizing the full value of Big Data analytics-related investments were (followed by the percentage of survey respondents that selected each respective choice):

1. Difficulty merging multiple, disparate data sources. (40%)
2. Lack of skilled Big Data practitioners. (39%)
3. Difficulty transforming data to suitable form for analysis. (37%)
4. Difficulty integrating Big Data with existing infrastructure. (37%)
5. Difficulty maintaining application performance for large volume of concurrent users. (35%)

The top five non-technology barriers to successfully realizing the full value of Big Data analytics-related investments were (followed by the percentage of survey respondents that selected each respective choice):

1. Difficulty "selling" the value to end-users. (41%)
2. Difficulty getting stakeholders to agreeing to data definitions. (38%)
3. Difficulty operationalizing insights. (36%)
4. Initial projects too ambitious. (35%)
5. Lack of data or inability to access data sources. (34%)

ANALYSIS: While these are high-level findings and much value can be attained by analyzing the results in more detail (which follows in this report), there are a number of takeaways from these findings.

First, it is clear that overcoming non-technology-related challenges are just as important to successful Big Data analytics projects as technology challenges. These include data governance, people and process-related issues. Specifically, getting stakeholders to embrace data-driven decision-making and abandon previous ways of thinking is a challenge that crosses vertical industries. This is largely an exercise in change management and can make or break Big Data analytics projects.

Second, the often-discussed Big Data skills shortage is real and is having a material impact on Big Data analytics projects. Based on qualitative research, namely discussions with Big Data analytics practitioners, it is clear that this skills shortage spans both IT-related skills, including Hadoop development and administration, and business-related functions, including data science.

Third, data itself presents challenges. Finding the right data sources, integrating multiple sources of data, and transforming data into forms suitable for analysis (i.e. applying some level of structure to unstructured data) are among the top barriers to successful Big Data analytics projects and, like other barriers to success identified, cross vertical industries and use cases.

CASE STUDY:

Seattle Children's Hospital is actually three institutions in one. The clinical hospital specializes in medical treatment of children from birth through young adulthood. The research arm's focus is pediatric medical research spread across specialties including cancer, genetics, immunology, pathology, infectious disease, injury prevention and bioethics. Seattle Children's foundation arm's mission is to raise funds to support both the clinical and research arms.

Eugene Kolker is Seattle Children's Chief Data Officer. He and his team focus on data analysis services to both improve patient outcomes and increase operational efficiencies. For example, Kolker is leveraging Big Data analytics to find ways to improve complex disease management for chronic conditions such as diabetes. His team also analyzes the hospital's use of various assets to find ways to improve resource allocation. The insights Kolker and his team generate are used by practitioners – both clinicians and management – throughout the hospital. (As complex as the data analysis is, Kolker highlights another challenge he and his team face: selling the value of data analysis to end-users. "You have to be a data storyteller. Otherwise you're going to be somewhere and people will be in another place." Kolker says to get end-users on board, it is important to include them in the development of Big Data analytics projects from the beginning. This way, end-users feel invested in projects and that they have a say in their outcomes. Otherwise, says Kolker, "you're not going to make a big impact."

Top Barriers by Phase of Deployment

Of course, there are many barriers faced by those adopting Big Data analytics. Some barriers are troublesome and consistent throughout the entire journey to adoption, while others are less troublesome overall. Moreover, some barriers are more or less challenging at different stages of the adoption process.

Wikibon had each survey respondent class themselves on the following scale:

1. We are currently evaluating Big Data analytics use cases and vendors/technology.
2. We have at least one Big Data analytics pilot/proof-of-concept project underway.
3. We have at least one Big Data analytics deployment in production supporting mission critical business processes and/or applications.
4. We have at least one Big Data analytics pilot/proof-of-concept project underway and at least one Big Data Analytics deployment in production supporting mission critical business processes and/or applications.

For our purposes, Wikibon combined the responses from those respondents classified in groups three and four into one class, that being those that have at least one Big Data analytics project in production.

Evaluation Phase, Top Barriers

Based on the survey results, for those Big Data analytics projects in the evaluation phase (n=123) the top three technology-related barriers to success (followed by the percentage of survey respondents that selected each respective choice):

1. Lack of skilled Big Data practitioners. (50%)
2. Difficulty integrating Big Data with existing infrastructure. (46%)
3. Confusion/uncertainty regarding the vendors/technologies to use. (42%)

The top three non-technology-related barriers to success for those Big Data analytics projects in the evaluation phase are (followed by the percentage of survey respondents that selected each respective choice):

1. Difficulty selling the value to end-users. (42%)
2. Lack of executive/management buy-in. (40%)
3. Lack of data or inability to access data sources. (38%)

ANALYSIS: During the evaluation phase of Big Data analytics projects, practitioners are evaluating not just the use cases, technologies and vendors, but also the internal skills required to put them into practice. It is not surprising, then, that a lack of skilled Big Data practitioners is one of the main areas of concern for those in the evaluation phase. And not only must practitioners identify skilled colleagues to take part in Big Data analytics projects, they must also convince them of the value of taking part, hence the top non-technology challenge of selling the value of Big Data analytics projects to end-users. Further, these practitioners must get buy-in from management, if for no other reason than to secure funding for projects.

At the evaluation stage, practitioners are very much trying to marshal the resources (funds, people and technologies) to execute a proof-of-concept, but also make the case to end-users and management that such projects will return valuable insights that will result in measurable improvements to the business (be it better efficiency, improved customer satisfaction or some other metric.) Clearly, a key skill for Big Data analytics practitioners taking the lead in such projects is communication. Without the communication skills to convey the potential value of Big Data analytics projects, getting new projects off the ground becomes difficult. Getting projects launched with a simple objective with the right level of visibility that shows the value should be the primary goal for practitioners at this stage.

Proof-of-Concept Phase, Top Barriers

Based on the survey results, for those Big Data analytics projects in the proof-of-concept phase (n=85) the top three technology-related barriers to success are (followed by the percentage of survey respondents that selected each respective choice):

1. Difficulty merging disparate data sets. (51%)
2. Technology too raw and difficult to use. (48%)
3. Lack of skilled Big Data practitioners. (47%)

The top three non-technology-related barriers to success for those Big Data analytics projects in the proof-of-concept phase are (followed by the percentage of survey respondents that selected each respective choice):

1. Selling value to end-users. (62%)
2. Difficulty getting stakeholders to agree. (49%)
3. Initial projects too ambitious. (46%)

ANALYSIS: Once funding is secured and technologies chosen, practitioners move from the evaluation stage to the proof-of-concept phase. Here the goal is to execute relatively simple analytics projects in ways that clearly show value and thus justification for moving to full-scale production. As such, the top technology-related challenges during the proof-of-concept phase, as identified by survey respondents, are difficulty merging disparate data sets, working with relatively immature or raw technology, and (as with the evaluation stage) finding enough skilled practitioners to execute the project. From a non-technology perspective, the biggest obstacles standing between practitioners and successful proof-of-concepts are selling the value of projects to end-users, getting stakeholders to agree on data definitions, and simply realizing initial projects were too ambitious from the get-go.

As discussed earlier in this report, most Big Data analytics practitioners have merged disparate data sets for analysis. Based on the survey findings, this is hardly a trivial exercise, but clearly an important one. Indeed, some argue that if you are not blending data sources for analysis, you aren't practicing Big Data analytics. The challenge lies in various structures of disparate data sets (or lack of structure, in some cases) and blending data sets of differing structure is significantly more complex than traditional data management. It is during proof-of-concepts that practitioners first run into this reality. That the technology itself is difficult to use, including data integration, doesn't help matters.

Finding skilled practitioners to take part in Big Data analytics projects is still a challenge in the proof-of-concept phase, but we also see the emergence of data governance challenges here. Getting stakeholders to agree on data definitions is a notoriously thorny problem even in "traditional" business intelligence environments. What constitutes a customer, for example, can vary by department or even by groups within departments. This problem is exacerbated in Big Data analytics projects when new data sources are being introduced and require consensus on how to classify and tag resulting entities.

Finally, it is important for Big Data analytics practitioners to start with manageable projects with realistic objectives. Practitioners that attempt to change enterprise-wide business processes and deploy a "comprehensive" Big Data architecture/platform in initial projects are setting themselves up for failure.

Production Phase, Top Barriers

For those Big Data analytics projects in production (n=95) the top three technology-related barriers to success are (followed by the percentage of survey respondents that selected each respective choice):

1. Difficulty merging disparate data sets. (45%)
2. Difficulty transforming data into suitable form for analysis. (44%)
3. Difficulty maintaining application performance as data volumes increases. (38%)

The top three non-technology-related barriers to success for those Big Data analytics projects in production are (followed by the percentage of survey respondents that selected each respective choice):

1. Difficulty operationalizing insights. (41%)
2. Lack of data or inability to access data. (40%)
3. Poor data or information quality. (38%)

ANALYSIS: Once proof-of-concepts are completed successfully, there are still plenty of challenges left. Practitioners involved in production Big Data analytics deployments continue to struggle with merging disparate data sets as well as transforming the resulting data into forms suitable for analysis. These challenges become more difficult in production environments when data volumes increase dramatically and the pace at which data must be integrated and transformed likewise speeds up considerably. Maintaining application performance, which is a non-negotiable requirement for mission-critical applications, at scale is an ongoing challenge.

Other obstacles that emerge at this stage in the Big Data analytics project lifecycle include the ability to operationalize the insights mission-critical applications are generating. It is one thing to identify an insight that could have a material impact on the business, but integrating those insights and corresponding actions into complex and often entrenched business processes is another. Practitioners would be wise to start discussing how Big Data analytics-driven insights can be realistically integrated into the way the enterprise operates before reaching the production stage.

In production phase, data governance-related challenges take on added significance. Gaining access to the right data sources, while sometimes a technology-related issue (i.e. getting systems that speak different languages to talk to one another), more often the barriers are people-related. Business users and technology professionals alike tend to cling to their data sets and often bristle at the idea of sharing them with the wider enterprise. Again, communication and interpersonal skills are critical here to get all stakeholders on board and moving towards a common objective in order to reduce the impulse to hoard data.

Top Barriers by Role

There are a number of different roles associated with Big Data analytics projects and each see the challenges of Big Data analytics from different perspectives and lenses. It is important for practitioners to understand the common challenges they are likely to run up against in the context of their function supporting Big Data analytics projects. Likewise, vendors need to understand the concerns of those in various roles in the adoption process.

To review, Wikibon asked survey respondents to self-identify their primary role in regards to Big Data analytics projects based on the following classifications:

- Infrastructure administrator (i.e. A datacenter professional who manages infrastructure and hardware associated with Hadoop, NoSQL database and other technologies that support Big Data analytics projects.)
- Data scientist (i.e. An advanced analytics professional who conducts sophisticated analytics and develops predictive models/algorithms on large volumes of "messy" Big Data.)
- Application developer (i.e. A developer who builds applications that leverage Big Data analytics such as predictive models and algorithms.)
- Business analyst (i.e. A departmental power-user who conducts analysis of various Big Data sets with tools such as Excel and SPSS.)
- Business user (i.e. A line-of-business professional who uses dashboards and other visualizations to understand Big Data).

Following is a breakdown of the top technology and non-technology barriers to successful Big Data analytics projects by role.

Infrastructure Administrators, Top Barriers

For Infrastructure Administrators (n=58), the top three technology-related barriers to successful Big Data analytics projects are (followed by the percentage of survey respondents that selected each respective choice):

1. Difficulty maintaining application performance as data volumes increase. (40%)
2. Difficulty transforming data into a suitable form for analysis. (38%)
3. Big data technology too raw and difficult to use. (38%)

For Infrastructure Administrators, the top three non-technology-related barriers to successful Big Data analytics projects are (followed by the percentage of survey respondents that selected each respective choice):

1. Difficulty getting stakeholders to agreeing to data definitions. (43%)

2. Difficulty "selling" the value to end-users. (40%)
3. Initial projects too ambitious. (38%)

Data Scientists, Top Barriers

For Data Scientists (n=42), the top three technology-related barriers to successful Big Data analytics projects are (followed by the percentage of survey respondents that selected each respective choice):

1. Big data technology too raw and difficult to use. (48%)
2. Difficulty integrating Big Data with existing infrastructure. (45%)
3. Difficulty maintaining application performance for large volume of concurrent users. (45%)

For Data Scientists, the top three non-technology-related barriers to successful Big Data analytics projects are (followed by the percentage of survey respondents that selected each respective choice):

1. Difficulty getting stakeholders to agreeing to data definitions. (55%)
2. Difficulty "selling" the value to end-users. (52%)
3. Difficulty operationalizing insights. (50%)

Application Developers, Top Barriers

For Application Developers (n=52), the top three technology-related barriers to successful Big Data analytics projects are (followed by the percentage of survey respondents that selected each respective choice):

1. Difficulty transforming data into a suitable form for analysis. (46%)
2. Technology lacks enterprise-grade back-up/recovery. (46%)
3. Technology lacks enterprise-grade security capabilities. (44%)

For Application Developers, the top three non-technology-related barriers to successful Big Data analytics projects are (followed by the percentage of survey respondents that selected each respective choice):

1. Unsure of regulatory/compliance/customer privacy implications. (42%)
2. Lack of executive/management buy-in. (42%)
3. Difficulty "selling" the value to end-users. (40%)

Business Analysts, Top Barriers

For Business Analysts (n=53), the top three technology-related barriers to successful Big Data analytics projects are (followed by the percentage of survey respondents that selected each respective choice):

1. Lack of skilled Big Data practitioners. (58%)
2. Confusion/uncertainty regarding the vendors/technologies to use. (43%)
3. Difficulty merging multiple, disparate data sources. (42%)

For Business Analysts, the top three non-technology-related barriers to successful Big Data analytics projects are (followed by the percentage of survey respondents that selected each respective choice):

1. Poor data and information quality. (40%)
2. Lack of data or inability to access data sources. (40%)
3. Lack of executive/management buy-in. (38%)

Business Users, Top Barriers

For Business Users (n=98), the top three technology-related barriers to successful Big Data analytics projects are (followed by the percentage of survey respondents that selected each respective choice):

1. Difficulty merging multiple, disparate data sources. (49%)
2. Big data technology too raw and difficult to use. (40%)
3. Confusion/uncertainty regarding the vendors/technologies to use. (38%)

For Business Users, the top three non-technology-related barriers to successful Big Data analytics projects are (followed by the percentage of survey respondents that selected each respective choice):

1. Difficulty getting stakeholders to agreeing to data definitions. (42%)
2. Difficulty operationalizing insights. (41%)
3. Lack of data or inability to access data sources. (40%)

ANALYSIS: The breakdown of the various challenges associated with realizing successful outcomes for Big Data analytics projects by role is telling. Infrastructure administrators are tasked with standing up the technology that underpins projects. This includes the hardware to support Hadoop, NoSQL and other distributed database clusters, as well as the networking gear to move data between systems and related storage platforms. They are also tasked with maintain systems running to support Big Data analytics workloads, hence the biggest technology-related challenge

of this group is maintaining performance at scale. Doing so is made more difficult by the relative immaturity of the related technology. Based on the responses of this group, infrastructure professionals are also helping their fellow practitioners with other parts of Big Data analytics projects, including data transformation and the all-important job of selling the value of Big Data analytics to the rest of the enterprise.

As for Data Scientists, the "rock stars" of Big Data analytics, the biggest barrier to successful projects is the immaturity and rawness of the related technology. This is an interesting result, as we often hear that the biggest bottleneck for Data Scientists, who sift through vast amounts of multi-structured data to find unlikely but valuable insights, is data transformation. While no doubt data transformation does account for a large part of data science work, our survey results indicate that the tools at their disposal are another significant barrier for Data Scientists. Other challenges for this group include getting stakeholders to agree on data definitions and selling the value of Big Data analytics to end-users. Both of these non-technology-related barriers continue to arise across roles and phase of deployment.

Application developers have a difficult job. They are tasked with operationalizing the insights generate by Data Scientists. This requires building applications that leverage advanced algorithms run against large volumes of high velocity data. A key issue when operationalizing Big Data analytics for developers is understanding the implications of their work. As has been said, just because Big Data makes something possible doesn't mean you should do it. As such, uncertainty around regulatory, compliance and privacy issues is a top concern for application developers. Enterprise-grade back-up & recovery and security capabilities are top technology challenges for this group, and is related to the aforementioned compliance issues. Securing applications to stay on the right side of regulatory frameworks and even self-imposed corporate policies is top of mind for developers.

Business analysts are the power-users inside line-of-business departments. In many ways they serve as the liaison between the Data Science team and business users. They sometimes work with Data Scientists but just as often work within their business group to produce key data and insights used to make both strategic and tactical decisions. Many business analysts are Excel experts, but are increasingly looking to emerging Big Data analytics technologies to support their work. Unfortunately, what they often find leaves them scratching their heads thanks to the relative immaturity of the technologies and "Big Data washing" by vendors. Data and information quality is also a major concern of this group, not surprising since their analysis is often the basis for day-to-day operational decisions.

Finally, business users are those line-of-business managers and other executives that consume Big Data analytics, usually via data visualization and business intelligence tools. Business users, ultimately, are those that make the important strategic decisions that impact the larger enterprise as well as daily tactical decisions that keep the business humming. Based on the survey results, when it comes to Big Data analytics technologies to support data-driven decision-making, business users find the technology difficult to use and the vendor marketing confusing. But they are not just passive consumers of insights. Business users identified difficulty merging multiple data sets as a top technology-related challenge, meaning they are also analytics practitioners. Data governance is also a challenge for business users, as getting their colleagues to agree to uniform data definitions is among the top non-technology barriers to Big Data analytics success for this group.

Hadoop in the Enterprise

Considering its important role as a foundational Big Data analytics technology, Wikibon believes it is worth drilling into the state of Big Data analytics projects that involve Hadoop as well as identify the main barriers to success for these projects in particular.

The State of Hadoop in the Enterprise

As part of the survey, we asked just those respondents that said they have already deployed Hadoop (36%) about their use and planned use of the open source Big Data framework.

The top use cases for Hadoop among respondents (n=110) are (followed by the percentage of survey respondents that selected each respective choice):

1. IT equipment support. (54%)
2. Data transformations. (52%)
3. Risk management. (51%)
4. Product development support. (48%)
5. Network analysis. (45%)

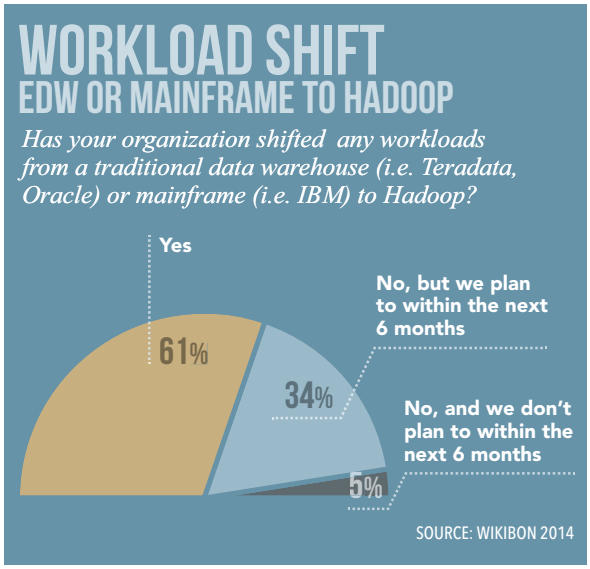

FIGURE 12

One topic getting significant attention from vendors and practitioners alike is the relationship between Hadoop and the enterprise data warehouse. As highlghted earlier in this report, a significant proportion of practitioners consider the data warehouse an important part of Big Data Analytics projects. Still, of those responents using Hadoop, 61% said they had shifted at least one workload from a data warehouse or mainframe to Hadoop. Another 34% siad they plan to shift workloads in the future **[Figure 12]**.

A number of Big Data analytics practitioner, during in-depth conversations with Wikibon, specifically said they expect to continue shifting workloads from data warehouses to Hadoop as the latter adds features and functionality. In some cases, these preactitioners believe their enterprises' data warehouse spend will significantly slow and potentially freeze in the coming years.

We also asked Hadoop practitioners about how they sourced the technology. Only 25% of Hadoop practitioners are paying customers of one or another Hadoop vendor. 24% use a free distribution provided by a vendor, but the majority, 51%, roll their own Hadoop downloaded from the Apache Software Foundation **[Figure 13]**.

CASE STUDY:

The OCLC's is a non-profit library cooperative whose mission is to help its member libraries better organize and share the world's information. That means the OCLC deals with lots of data. One of its services is the WorldCat database. WorldCat stores bibliographic data and information on books, CDs, research articles and other library content and services that is shared by OCLC's member libraries. WorldCat is made available to anyone on the web, where people can search the database, using any number of criteria, in order to locate books, articles and other resources. To date, over 72,000 libraries worldwide have contributed bibliographic data representing over two billion items (books, CDs, articles, etc.) to WorldCat.

Until recently, OCLC used a traditional relational database to support WorldCat. But with the explosion of data showing no signs of slowing, the organization decided it needed to move to a more sustainable approach. One issue was storage costs. Each time OCLC needed to expand its RDBMS cluster due to growing data volumes, the organization was looking at significant costs both for hardware and software licensing. Another issue was performance. The RDBMS was taking too long run some operations. One workload in particular, which processed the entire WorldCat dataset to more intelligently group similar records, was going to take three months to complete.

OCLC decided to shift these workloads from its existing RDBMS to Hadoop. Specifically, OCLC deployed an HBase cluster based on Cloudera's Hadoop distribution and shifted support for WorldCat – storage, data processing and search capabilities – to the Hadoop-based database. The results were significant. Storage costs plummeted, according to Ron Buckley, Senior OCLC Technology Manager and leader of the Hadoop migration team. And performance improved, in some cases dramatically. The aforementioned workload that would have taken three months to run on OCLC's previous RDBMS cluster? Now that process runs every morning before I get here," says Buckley. The significance of these improvements is that OCLC is now in a position to continue supporting its members with these and new Big Data services well into the future.

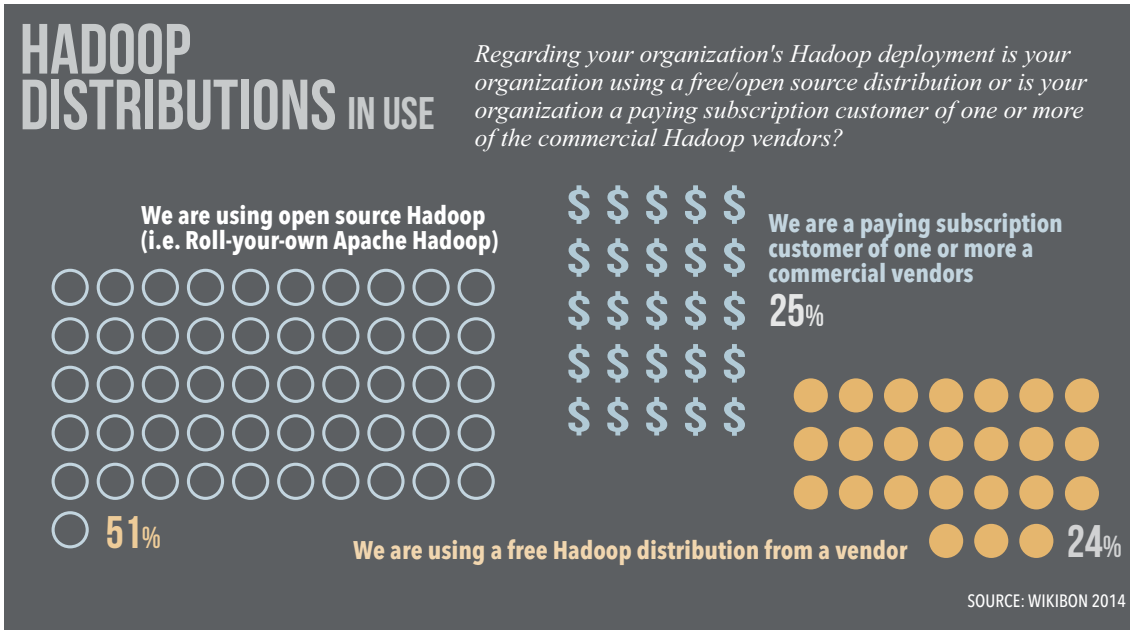

FIGURE 13

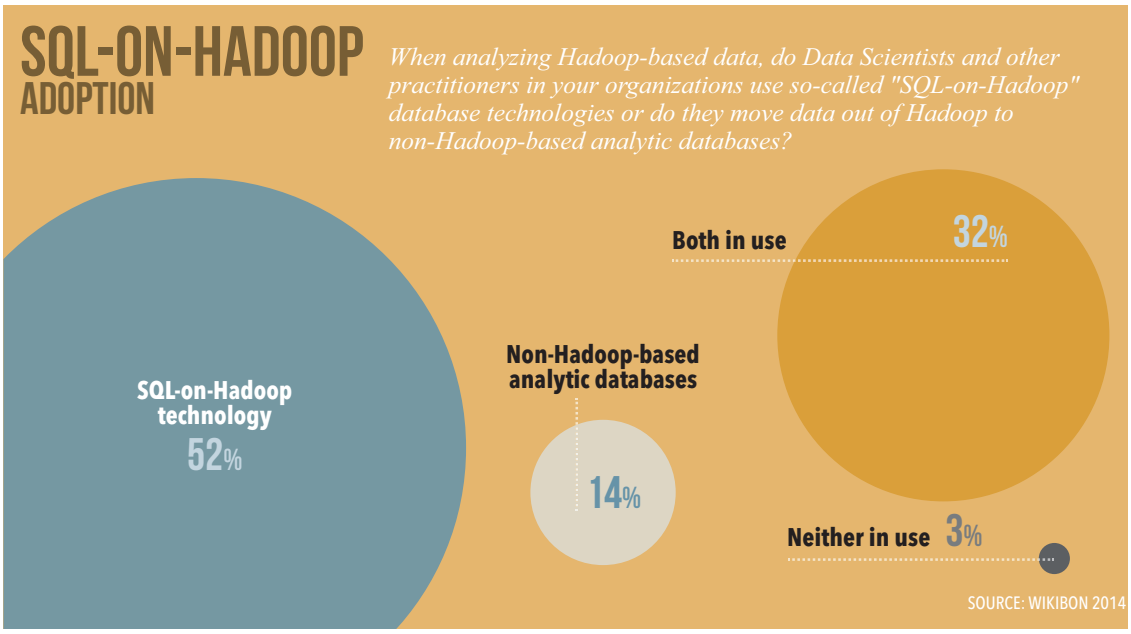

FIGURE 14

Another hot topic in the Hadoop community is SQL-on-Hadoop (or SQL-in-Hadoop) **[Figure 14]**. The basic premise behind SQL-on-Hadoop is to provide users the ability to manipulate Hadoop-based data with a language they now well (SQL) and without having to move data out of Hadoop for analysis.

While SQL-on-Hadoop tools are still raw, a majority of survey respondents (84%) reported the use of the technology by Data Scientists and analysts working with Hadoop. A sub-segment of these respondents are using other methods in conjunction with SQL-on-Hadoop.

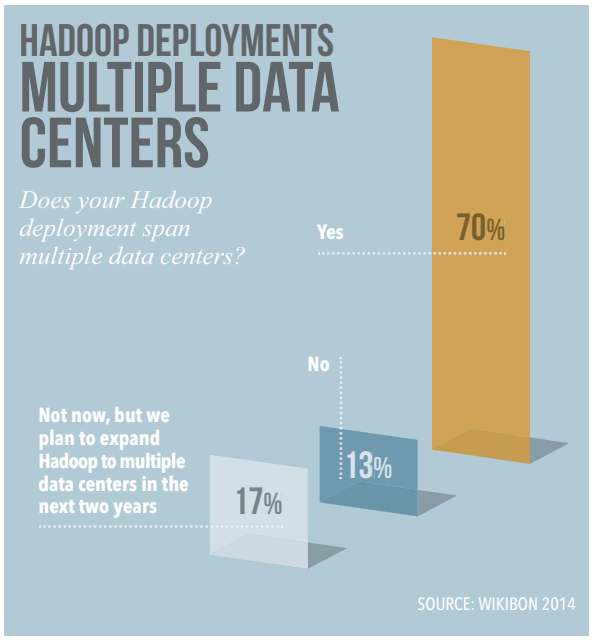

FIGURE 15

To get a sense of their commitment to Hadoop, we asked respondents if their Hadoop deployments spanned multiple data centers **[Figure 15]**. A staggering 70% reported multiple data center Hadoop deployments, with 17% planning to expand to multiple data centers in the future.

Top Barriers for Hadoop Projects

For Hadoop practitioners, the top three technology-related barriers to successful Big Data analytics projects are (followed by the percentage of survey respondents that selected each respective choice):

1. Lack of enterprise grade back up and recovery. (44%)
2. Lack of enterprise-grade security. (43%)
3. Difficulty merging disparate data sets. (41%)

For Hadoop practitioners, the top three non-technology-related barriers to successful Big Data analytics projects are (followed by the percentage of survey respondents that selected each respective choice):

1. Selling value to end-users. (44%)
2. Difficulty getting stakeholders to agree to data definitions. (42%)
3. Lack of executive buy-in. (38%)

The majority of respondents using Hadoop, 64%, are doing so in proof-of-concept environments. What we wanted to know is what, if anything, is preventing these projects from moving to full-scale production supporting mission-critical applications. The top barriers are (followed by the percentage of survey respondents that selected each respective choice):

1. Concerns about a lack of enterprise-grade backup and recovery in Hadoop. (53%)
2. Concerns about a lack of enterprise-grade high-availability in Hadoop. (48%)
3. Concerns about maintaining performance at scale in Hadoop. (45%)

ANALYSIS: Like Big Data analytics generally, Hadoop use cases run the gamut and are applicable across vertical industries. Notably, however, while many of these use cases represent net-new applications, over 60% of Hadoop practitioners have moved at least one existing application from a data warehouse or mainframe. This has significant implications for data warehouse and mainframe vendors, as this trend is being driven by the relative inflexibility of legacy technologies as compared to Hadoop and the large disparity in costs. Legacy data warehouse vendors, in particular, must find ways to adapt to this new reality. Hadoop is not going to replace the data warehouse, at least not any time soon. But the open source Big Data framework does pose a threat to the cushy profit margins of data warehouse vendors.

This trend is expected to continue as Hadoop vendors improve so-called SQL-on-Hadoop offerings. These technologies from vendors such as Cloudera, Hortonworks and Pivotal allow data scientists and less sophisticated business analysts and business users to query data in Hadoop using ANSI SQL-like tools (some SQL-on-Hadoop offerings have more complete SQL functionality than others) and developers to build Hadoop-based data-driven applications.

Another striking finding is that just a quarter of Hadoop practitioners are paying customers of one commercial Hadoop vendor or another. This finding indicates that these practitioners believe they are sophisticated enough to make use of Hadoop without vendor support, at least initially. As proof-of-concepts move to production, however, these findings suggest there is a major opportunity for vendors to provide enterprise-level support services. Based on in-depth discussions with Hadoop practitioners, most believe they will look to vendor support once projects move to production if for no other reason than corporate policies require it. But vendor support is also important once Hadoop and related SQL-on-Hadoop tools are in wide use by Data Scientists, business analysts and business users.

As for the top barriers to succesful Hadoop-related Big Data analytics projects, enterprise-grade features top the list of technology-related challenges. Before practitioners are willing to rely on Hadoop

to support mission-critical applications, they want reassurance that the open source framework can meet the necessary back-up & recovery and security requirements of the enterprise. While this is a sign to the Hadoop community and the vendor community to make haste in these areas, it is also a sign that many practitioners have moved beyond concerns about Hadoop's analytics functionality.

Data governance is also a concern for Hadoop practitioners. These concerns are not unique to Hadoop, but these survey results should reinforce the notion that data governance – data retention policies, data access policies, data quality procedures – cannot be ignored when it comes to Hadoop, either by practitioners or vendors.

Conclusions

It is clear from the survey results that practitioners believe Big Data analytics holds great promise, but most are early on in the journey. Those that have taken the first steps to evaluate Big Data analytics projects and embark on proofs-of-concept projects as well as those that have deployed Big Data analytics in production offer valuable lessons to their colleagues that will soon follow suit.

ACTION ITEM: Big Data analytics practitioners must consider the specific challenges associated with each phase of the project life-cycles as well as pay particular attention to those obstacles most closely associated with their individual roles in projects. While some challenges span phases and roles (most notably the challenge of selling the value of Big Data analytics to end-users), others are unique to particular points in the project life cycle and/or to particular roles/functions.

Big Data analytics vendors likewise need to do more to understand the particular circumstances of projects in which they are engaged. The requirements that must be met, as discussed throughout this report, vary by project, phase of deployment and role or function within projects. Success is largely dependent on meeting these requirements at the right point and with the right participants.

ABOUT WIKIBON

Wikibon is a research and advisory firm focused on disruptive technologies and their impact on the enterprise. Wikibon's core coverage areas are Big Data, cloud computing and software-led infrastructure. Wikibon was the first research firm to size and forecast the Big Data market in 2011, and its analysis of the evolving Big Data ecosystem and technology landscape is relied on by practitioners, vendors, venture capitalists and more. For more Wikibon Big Data research and analysis, visit Wikibon.org/BigData. To become a Wikibon client, please contact John Greco at 774-463-3400 or john.greco@wikibon.org.