Operationalizing the Buzz: Big Data 2013

An ENTERPRISE MANAGEMENT ASSOCIATES® (EMA™) and 9sight Consulting Research Report

November 2013

Sponsord by:

IT & DATA MANAGEMENT RESEARCH, INDUSTRY ANALYSIS & CONSULTING

Table of Contents

Table of Contents (continued)

1. Executive Summary

The 2013 EMA/9sight Big Data research makes a clear case for the maturation of Big Data as a critical approach for innovative companies. This year's survey went beyond simple questions of strategy, adoption and use to explore why and how companies are utilizing Big Data. This year's findings show an increased level of Big Data sophistication between 2012 and 2013 respondents. An improved understanding of the "domains of data" drives this increased sophistication and maturity. Highly developed use of **Process-mediated**, **Machine-generated** and **Human-sourced** information is prevalent throughout this year's study.

The 2013 study dives deep into the Big Data project initiatives of EMA/9sight respondents focusing on multiple characteristics within each. These 259 respondents, averaging between two and three projects in their Big Data programs, provided information on nearly 600 ongoing Big Data efforts. Over 50% of these projects have an implementation stage of **In Operation** – **In Production** or **Implemented as a Pilot**. Respondents indicated that the top three business challenges were associated with **Risk Management** activities, **Ad-Hoc Operational** queries, and **Asset Optimization** operations. These projects provide groundbreaking detail information into not just the strategy of Big Data implementations, but also the details on implementation choices: on-premises vs. cloud; project sponsors throughout the organization specifically outside the office of the CIO; and actual implementation stages.

Speed of Processing Response has replaced **Online Archiving** as the top Big Data use case in the 2013 study. This shows that organizational strategies are moving from discovering "*the things we don't know we don't know*" into managing Big Data initiatives toward achievable business objectives and "*the things we know we don't know.*" That being said, many of the individual projects being implemented are still using an **Online Archiving** use case. **Speed of Processing Response** and **Online Archiving** are the two most popular uses cases in projects classified as **In Operation** indicating that these use cases are critical to early Big Data adopters.

Respondents in the 2013 survey indicated that the information consumers (users) of these Big Data projects are coming from the less technical ranks of their companies. Approximately 50% of users were from business backgrounds with **Line of Business Executives** and **Business Analysts** representing the top two responses. This shows that Big Data projects are moving beyond **Data Scientist** as the primary user of these projects. When examining the sponsors of Big Data projects, business is not only using the information results from these systems, but also "putting their money where their users are." Nearly 50% of all Big Data projects are sponsored by business organizations such as finance, marketing and sales. Just over two of ten Big Data projects were sponsored directly by the CIO.

Integrating Big Data initiatives into the fabric of everyday business operations is growing in importance. The types of projects being implemented overwhelmingly favor **Operational Analytics**. Operational Analytics workloads are the integration of advanced analytics such as customer segmentation, predictive analytics and graph analysis into operational workflows to provide real-time enhancements to business processes. An excellent example of Operational Analytics can be found as organizations move toward the real-time provisioning of goods and services. It is critical to provide visibility into AND action regarding illicit activities among customers. In addition, risk assessments become more important as businesses use value-based decisions to determine courses of action to pursue new customers and/or to retain existing ones.

The world of Big Data is maturing at a dramatic pace and supporting many of the project activities, information users and financial sponsors that were once the domain of traditional structured data management projects.

In summary, the world of Big Data is maturing at a dramatic pace and supporting many of the project activities, information users and financial sponsors that were once the domain of traditional structured data management projects. It is possible that within the next three to five years, Big Data will have fully absorbed those traditional approaches into a new world driven by a more open and dynamic set of data best practices.

1.1. Key Findings

The 2013 EMA/9sight Big Data research surveyed 259 business and technology stakeholders around the world. The survey instrument was designed to identify key trends surrounding the adoption, expectations and challenges associated with strategies, technologies and implementations of Big Data initiatives. The research identified the following highlights in the 2013 Big Data research and comparisons to the 2012 results:

- **Multiple projects within Big Data programs:** This year's respondents indicated that they had on average 2.5 projects in their Big Data Programs totaling 597 active Big Data projects.
- **Projects are In Operation:** Over 50% of these projects are **In Operation** – defined in this research as **In Production** or **Implemented as a Pilot**. This is a significant increase over 2012.
- **The Internet of Things is coming… if not here: Machine-generated** data represents the fastest growing data source for Big Data projects. This includes machine-to-machine and application log file information that contributes to linking devices to the Internet.
- **Big guys are getting into Big Data:** Enterprise-sized organizations made the largest jump in survey participation between 2012 and 2013. This indicates that Big Data programs are making their way into the most highly governed IT environment – the enterprise corporate data center.
- **Spreading around the globe:** Respondents in the Asia-Pacific (APAC) region showed the largest increase in response for the 2013 survey over 2012. Although the APAC region addresses Big Data with unique requirements, respondents provide insights into how Big Data is being utilized outside of North America.
- **Innovation knows no boundaries:** Over 70% of survey respondents who identified themselves as "innovators in Big Data" came from outside North America. The Europe-Middle East-Africa (EMEA) region was the single largest group with over 40% of Innovators.
- **Moving faster than ever before:** Of the Big Data Use Cases for our respondents, the top response was for **Speed of Processing Response** with over 50% of the total, illustrating that organizations are focusing less on exploring their data and more on how fast they can process information.
- **Corporate culture still matters:** Big Data is not just about technology. Some of the biggest obstacles to Big Data implementations relate directly to corporate culture. Issues of stakeholder communication and buy-in as well as coordinating implementation strategies are common challenges.
- **New brand of workload:** Operational Analytics – the integration of advanced analytics in realtime operational workflows – is the most prevalent type of project workload. From segmentation to asset optimization to risk management, Operational Analytics is pushing into critical business workflows.
- **Business is consuming Big Data information:** Nearly 50% of Big Data project users detailed in the 2013 study were business stakeholders: Line of Business Executives and Business Analysts from marketing, finance and customer care departments.

- **Economics are important:** Big Data technologies are applying pressure to the costs associated with many processing platforms. Top business challenges for 2013 respondents are **Improved Data Management, TCO** and **Improving Competitive Advantage.**
- **Big Data grows beyond the office of CIO:** Almost 50% of respondents indicated that funding for their Big Data initiatives originated from outside the overall IT budget. Finance, Marketing and Sales were the top non-CIO sponsors of Big Data projects.

2. Hybrid Data Ecosystem

In the 2012 "Big Data Comes of Age" study, EMA and 9sight identified that Big Data implementers and consumers are relying on a variety of platforms (not just Hadoop) to meet their Big Data requirements. EMA has established there is a collection of platforms that support Big Data initiatives. These platforms include new data management technologies such as Hadoop, MongoDB and Cassandra. But the collection also includes traditional SQL-based data management technologies supporting data warehouses and data marts; operational support systems such as Customer Relationship Management (CRM) and Enterprise Resource Planning (ERP); as well as cloud-based platforms leveraging freely available data sets from sources such as the Open Government Initiative (http://www.data.gov/) to Software-as-a-Service (SaaS) platforms such as Salesforce.com. EMA refers to this collection of platforms as the **Hybrid Data Ecosystem**. These platforms include:

- Enterprise or federated data warehouse
- Data marts
- Operational data stores
- Analytical database platforms/appliances
- NoSQL data store platforms
- Data Discovery platforms
- Cloud-based data solutions
- Hadoop and its subprojects

Each of the platforms within the **Hybrid Data Ecosystem** supports a particular combination of business requirements and processing challenges. This is a relatively unique approach when compared to traditional best practices. Rather than maintaining a single data store that supports all business and technical requirements at the center of this architecture, the **Hybrid Data Ecosystem** seeks to find the best platform for a particular set of requirements and link those platforms together.

2.1. Platform Trends

There were changes in the choices of EMA/9sight panel respondents concerning technology platforms from 2012 to 2013. The most significant of these differences between the 2012 and 2013 surveys focus on two platform types in particular: **Analytical Data Platforms/Appliances** and **Operational Data Stores**.

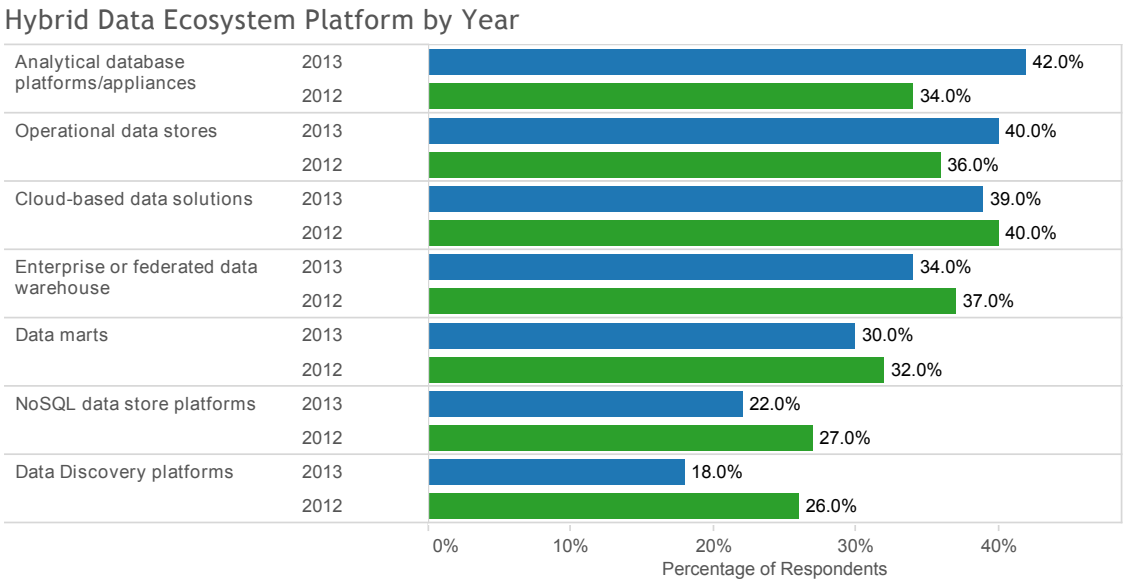


Figure 1

Analytical Data Platforms/Appliances made the largest jump in utilization, from 34% to 42% of respondents. This change reflects how important **Speed of Processing Response** is in Big Data use cases and the implementation of real-time **Operational Analytical** workloads. This also matches the workload types that **Analytical Data Platforms/Appliances** were designed to handle. The increase in responses for **Operational Data Stores** shows how Big Data initiatives are continuing to press into the everyday processes of organizations. From specific Big Data systems that handle order processing and point of sales to the inclusion of operational datasets into exploratory and analytical strategies, **Operational Data Stores** are some of the best sources of data to drive improvement in business processes, and by extension, competitive advantage.

Of the platforms that showed a decrease between 2012 and 2013, **NoSQL Data Stores** and **Data Discovery Platforms** fell to the last two places on the trend analysis. One of the main differences between the 2012 and 2013 surveys was the specific inclusion of Hadoop as a platform type separate from NoSQL Data Stores. This adjustment to the survey options also contributed to the drop in **Data Discovery Platforms**. Hadoop and Hadoop HDFS are considered components of many **Data Discovery Platforms** that bridge the gap between NoSQL and SQL access layers.

2.2. Ecosystem Diversity

When asked how many platforms were part of their Big Data initiatives, the EMA/9sight respondents indicated that a wide number of **Hybrid Data Ecosystem** platforms were important to their Big Data environments. The most common environment was **Two Platforms** with over 30% of responses.

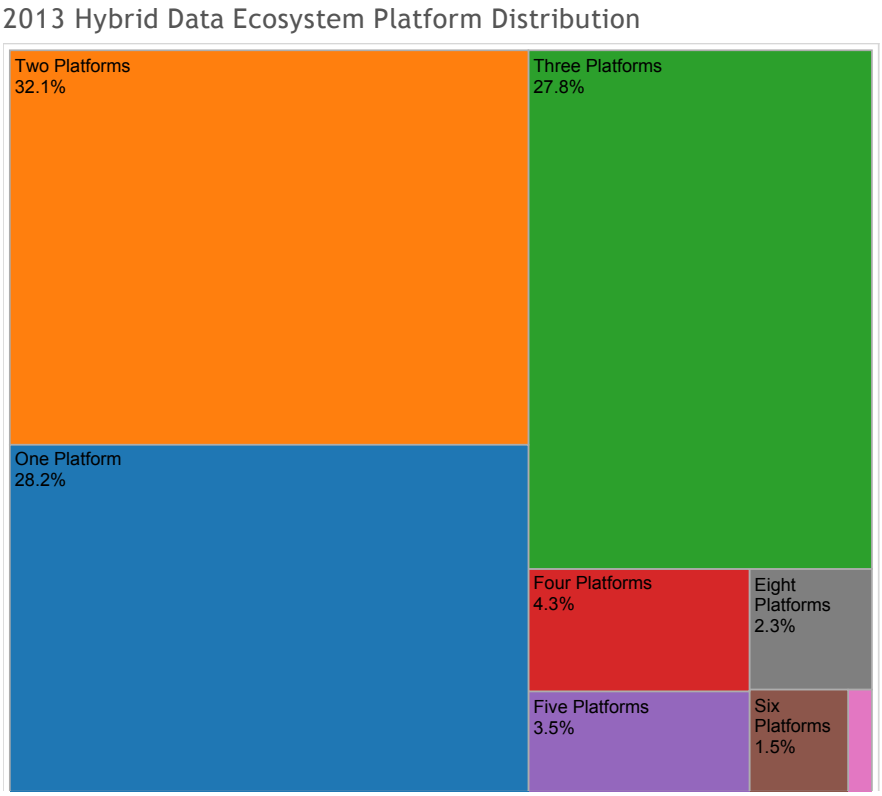


Figure 2

Nearly 65% of respondents are using two to four platforms, which indicates that they are implementing fairly complex and diverse combinations of technology to power their **Hybrid Data Ecosystem** environments.

2.3. Updates to the Ecosystem in 2013

For 2013, EMA expanded the definition of the **Hybrid Data Ecosystem** to include **Information and Data Management** and a focus on **Information Consumers**. Our 2013 results have also provided deeper insights into the workloads of this environment.

EMA expanded the definition of the Hybrid Data Ecosystem to include Information and Data Management and a focus on Information Consumers.

- ***Information and data management:*** *The 2012 research defined the number of platforms companies were using as well as how the platforms were related. In 2013, respondents provided deeper insights into how they choose to move information in a bi-directional manner between platforms and which technologies make that information management a reality.*
- ***Workloads:*** *The concepts of* ***Speed of Response*** *and* ***Complex Workload*** *were established in 2012 as key components of the* ***Hybrid Data Ecosystem*** *requirements. This year's research leveraged new project-based results to identify the workloads that Big Data initiatives are tackling. They included: Operational workloads associated with ordering, provisioning and billing for goods and services; Analytics workloads for summarizing, predicting and categorizing business operations; Operational Analytics workloads for the integration of analytical models into real-time business processes; and Exploration workloads designed to quickly and iteratively determine new uses for Big Data sources.*

- ***Information consumers:*** *In 2013, the role of information consumer or user was added to the Hybrid Data Ecosystem framework. As important as the underlying technology and processing results are, the users are the most important aspect of a Big Data initiative. Users are the direct links to the top and bottom line of the balance sheet and the best way to gauge the success or failure of a Big Data initiative.*

The following details the 2013 EMA Hybrid Data Ecosystem, supported by two years of extensive user research on Big Data initiatives.

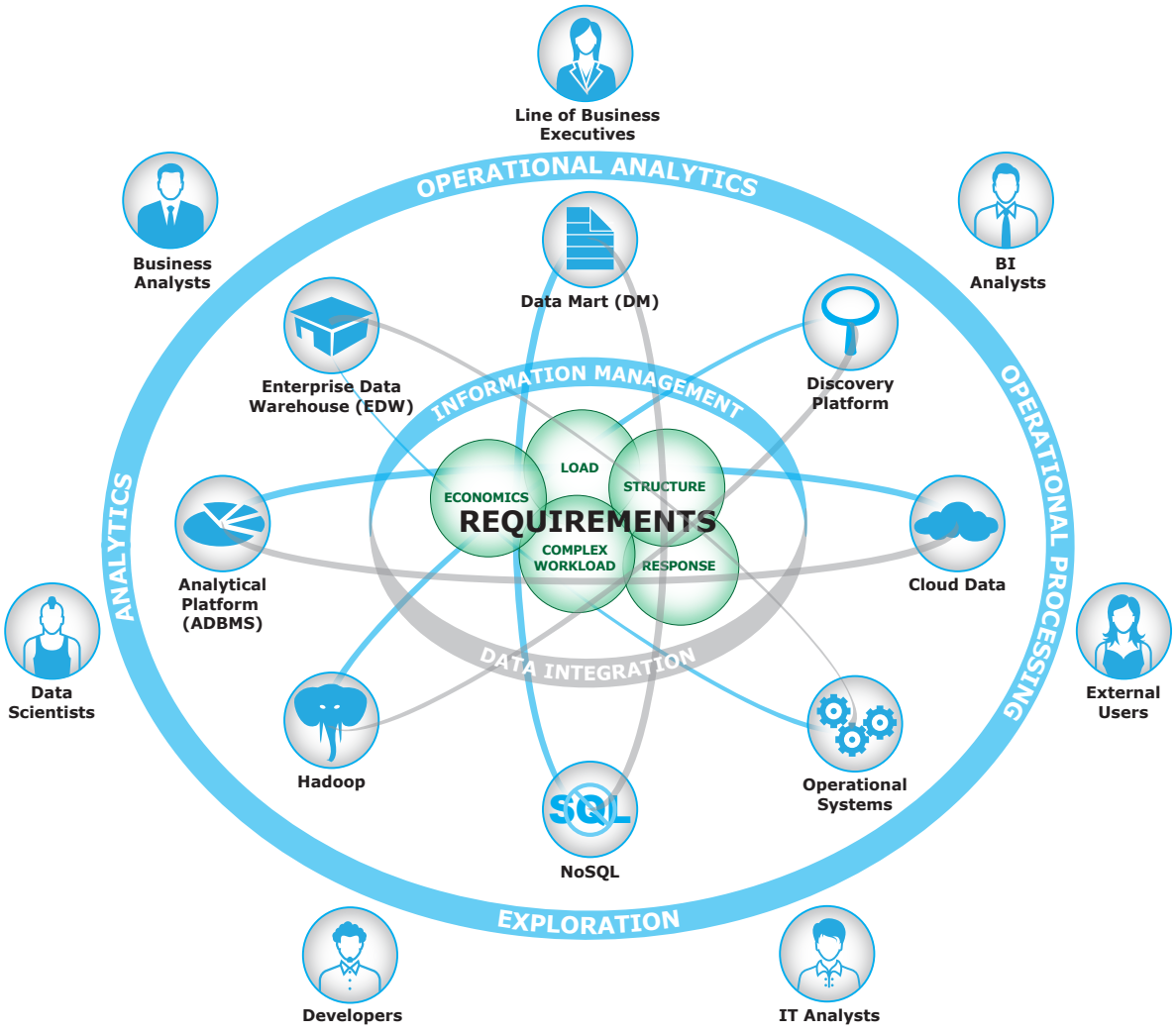


Figure 3: Hybrid Data Ecosystem

3. Operationalizing the Big Data Buzz

Since the publication of the inaugural EMA/9sight Big Data survey in November 2012, the marketing hype surrounding the topic has quieted considerably. Hype has a natural half-life; the chattering of the Big Data Geiger counter has slowed significantly. Interest has moved to Big Analytics. Of course, the two are related. The latter emerges from the former. Big Analytics is one of the more significant business uses of Big Data. The respondents to the 2013 survey confirm the trend; they have also shifted their attention from worrying about collecting and managing Big Data to using it and getting value from their investment. Contrary to suggestions by other observers, the respondents to the EMA/9sight survey have moved significantly from investigation and planning to operation between 2012 and 2013.

The respondents to the 2013 survey confirm the trend; they have also shifted their attention from worrying about collecting and managing Big Data to using it and getting value from their investment.

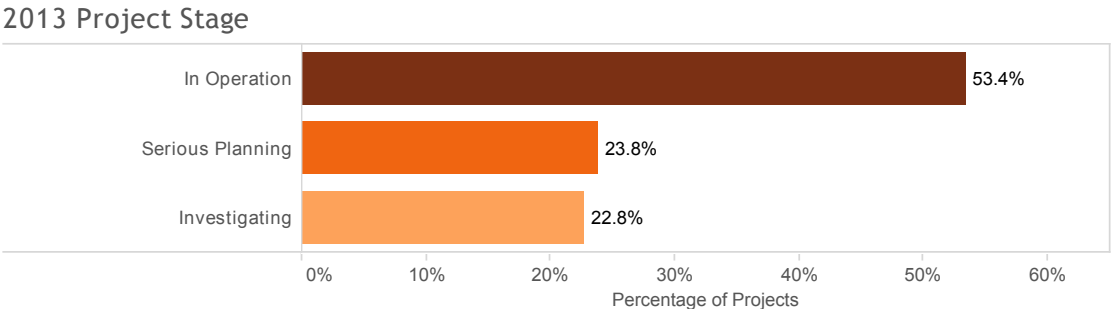


Figure 4

In 2013, 53% of respondents had an **In Operation** stage project, compared to 2012, when little more than a third of respondents indicated they had some aspect of their Big Data initiative **In Operation**.

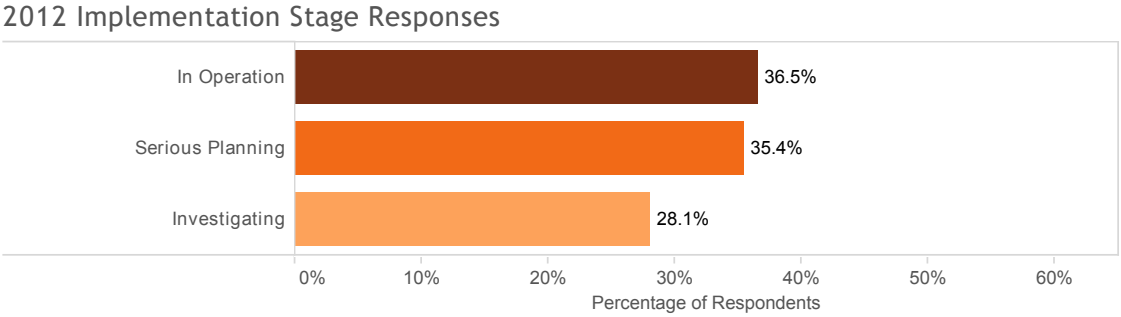


Figure 5

In the absence of a useful definition of Big Data and as a result of the findings, EMA/9sight declared in 2012 – somewhat riskily – that Big Data was simply all data. No better definition has appeared in the interim. This year's survey continues to show that businesses regard projects using traditional large data sets, such as Call Detail Records (CDRs) in Telecommunications or Point Of Sale (POS) receipts in the retail industry as fair game for processing with so-called Big Data tools and thus declaring them Big Data projects. EMA/9sight suspect that this may be politically motivated; Big Data projects may have a greater chance of approval simply because of the popularity of the term within an organization.

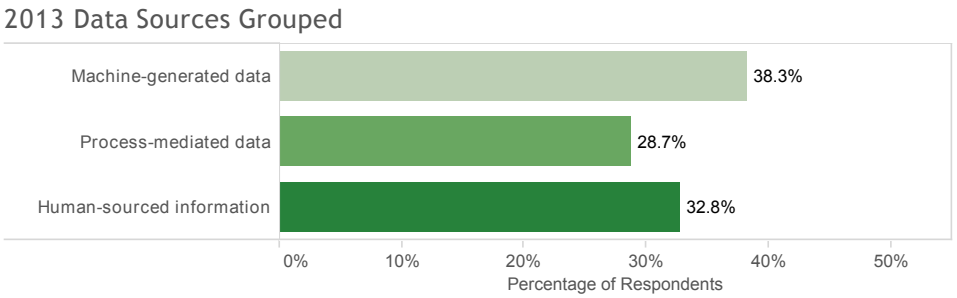


Figure 6

The results of the 2013 survey also show an increasing use of **Human-sourced** information and **Machine-generated** data in support of business goals.

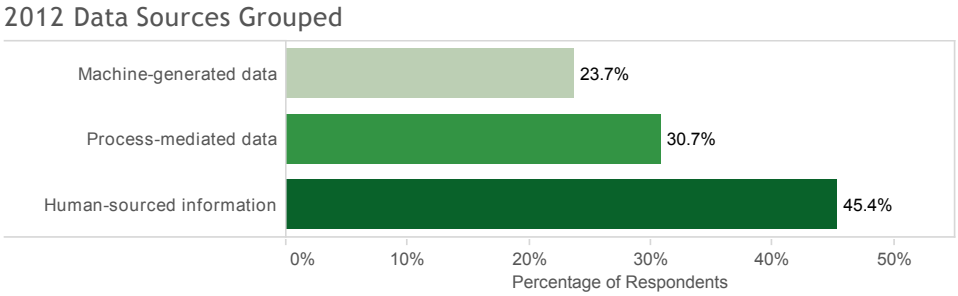


Figure 7

EMA/9sight projected in the 2012 results that many respondents had more than one Big Data project in progress. For this year's survey, EMA/9sight examined these projects individually where possible. The EMA/9sight panel respondents detailed nearly 600 unique Big Data projects.

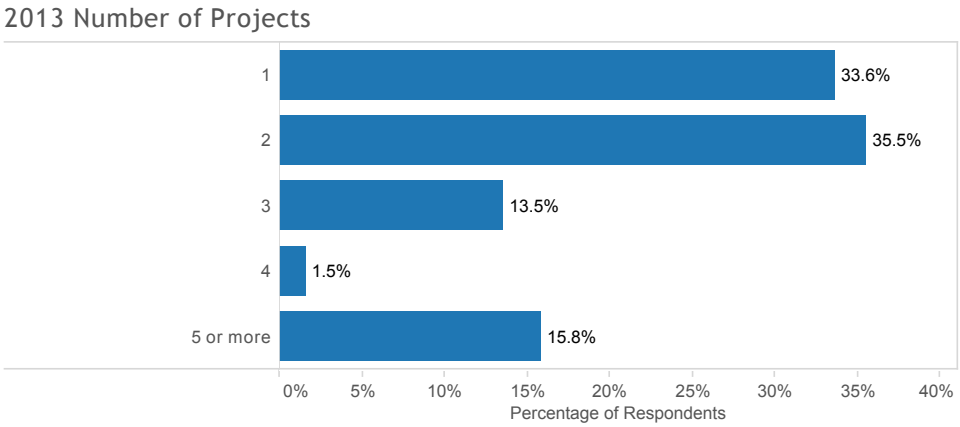


Figure 8

Approximately one third (33.6%) of respondents reported having one project as part of their Big Data program. Over one third (35.5%) had two projects. The remaining third had three or more projects as part of their Big Data initiatives. In fact, nearly 16% of respondents reported having five or more projects ongoing. These project numbers strongly suggest that Big Data is now deeply embedded in many companies.

3.1. The Evolving Big Data Tool Set

The hype may have abated from the vendors in the past twelve months, but new and improved function is still being delivered.

According to Wikibon[1], the total Big Data market, including hardware, software and services, stood at over $11 billion in 2012. In 2013, it is projected to reach some $18 billion – an annual growth of just over 60%. According to the author, "This puts it on pace to exceed $47 billion by 2017. That translates to a 31% compound annual growth rate over the five year period 2012–2017." Unfortunately, the count does include the rebranding of existing products as Big Data approaches, and no consideration is given to the consequent decrease in other markets, such as BI (Business Intelligence), in which these rebadged tools previously resided. This implies a considerably lower actual overall growth rate.

The long-awaited Hadoop 2.0, which shipped in October 2013, introduces HDFS Federation – removing a long-standing bottleneck and single point of failure in the Hadoop Distributed File System[2] (HDFS), and YARN[3] – splitting resource management and job life-cycle management into separate components. Overall, the effect is to allow Hadoop to move beyond its batch heritage and begin to handle more real-time work. Of course, this is similar to the transition that happened decades ago with the introduction of transactions managers such as CICS as well as hierarchical, and later, relational databases. The result of these introductions will be an environment more suited to mainstream enterprise needs. Nonetheless, being an open source environment, there is no single right answer. Many vendors are pursuing similar goals to enable the sort of real-time SQL interaction that many users and applications demand. How these new and existing components coexist and interoperate remains to be seen.

The result of these introductions will be an environment more suited to mainstream enterprise needs.

The continuing growth in Hadoop distributions (distros) poses the same question. The plethora of distros available from both specialized players and large hardware/software vendors have seen regular updates in the past year, indicating a vibrant marketplace and offering a wide variety of choices to prospective implementers. This choice, nonetheless, continues to be a two-edged sword of immense proportions; every distro contains its own special addendums and favorite tools in addition to base components, all at varying release levels, raising the specters of incompatibilities and lock-in after the extensive evaluation is done and the initial choice made. While such challenges appear to be acceptable to the open-source community at large, they represent a return to the bad old days for database and BI departments who are used to a more tightly controlled and integrated environment.

Beyond Hadoop, NoSQL[4] data stores are growing in popularity. As seen at DB-Engines.com[5], NoSQL platforms took six (6) of the top 20 places in an October 2013 ranking. Of course, a popularity rating based on web mentions/searches and installation numbers are not the same thing, but NoSQL certainly continues to elicit interest. Early suggestions that NoSQL would replace traditional relational databases have, as expected, faded in the face of the reality that NoSQL is not a panacea for all ills – it has its own particular strengths and weaknesses. Furthermore, the market dominance of and ongoing upgrades to relational databases have enabled the existing vendors to embed NoSQL type function in existing products to address the needs of less demanding non-traditional applications that previously had a strong incentive to switch to NoSQL. Although there are numerous specific strengths of the wide range of specialized technologies included in the NoSQL area, a reasonable overview is that

[1] Kelly, Jeff, "Big Data Vendor Revenue and Market Forecast 2012-2017", Wikibon, Sept. 16, 2013, http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2012-2017
[2] http://wiki.apache.org/hadoop/HDFS
[3] http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html
[4] http://en.wikipedia.org/wiki/Nosql
[5] http://db-engines.com/en/ranking

its core application is in areas demanding flexibility and performance in handling large volumes of semi-structured data with an unpredictably changing structure. NoSQL can support both operational and limited informational needs in such circumstances, but cannot replace an EDW (Enterprise Data Warehouse) for data consistency and reconciliation.

The consolidation of the relational database (RDBMS) market continued, albeit more slowly in 2013. Of course, new start-ups continue to emerge, but the focus for now is on the big RDBMS vendors integrating and enhancing their original databases and more recent acquisitions to better handle bigger data and faster queries. EMA/9sight respondents continue to use traditional relational databases in their Big Data projects.

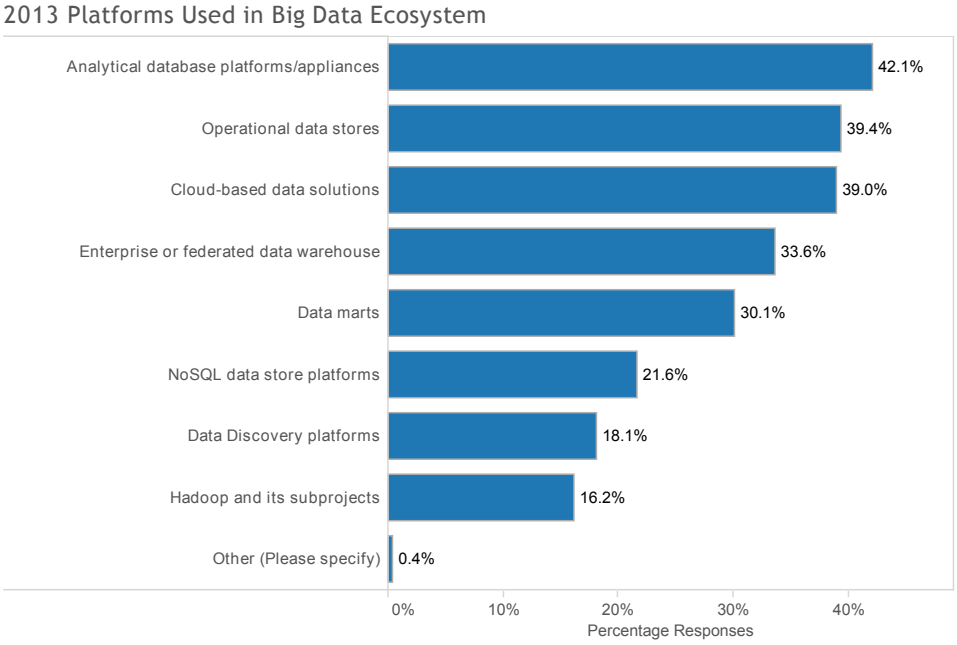


Figure 9

The other key trend, continuing from 2012, sees all the RDBMS and BI tool vendors enabling seamless access to the Hadoop HDFS from SQL-based tools via the Apache Hive[6] interface. Such an approach is well aligned with the Big Data architecture described below.

In addition, the emergence of in-memory and solid state[7] (SSD) storage support among data management platforms, as well as pervasive massively parallel processing is beginning to erode the boundary between operational systems, enterprise data warehouse and data marts, and moving all these systems toward a more real-time approach. This is a fundamental shift with wide-ranging consequences that will take a number of years to play out. In common parlance, this tends to be lumped with Big Data, but the reality is that they are two very different concepts. IT shops should take care to avoid overstretching themselves to try to address both areas at once.

3.2. Focusing on Real-time Business Value

The participants in the 2013 EMA/9sight survey indicated overwhelmingly that the business value of Big Data is to be found in its immediate use. Nearly half of the almost 600 projects were focused on the area of Operational Analytics. Add the further 20% that were designed for Operational Processing, and we see how real-time and near real-time use dominates Big Data. Similarly, the Use Case question

[6] http://hive.apache.org/
[7] http://en.wikipedia.org/wiki/Solid-state_drive

elicited the Speed of Processing response in over 50% of cases. This is unsurprising. Most Big Data is transitory in nature; its value has a very short shelf life. Sensors record instantaneous events or ever-changing measures. Social media reflects the volatile ebb and flow of personal opinion and social whims. All such data is of value in understanding the current moment in time, in the hope of predicting how these passing phantoms will drive behaviors that increase or decrease business value in legally binding transactions.

3.3. Defining a New Big Data Architecture

In the 2012 survey report, EMA/9sight defined the tri-domain information model. This model subdivides Big Data and, indeed, all data into three well-delimited categories based on characteristics of timeliness and flexibility that result from the sourcing and pre-processing of the data. The three domains identified are:

- ***Process-mediated data**, created by well-defined business processes and highly managed (typically) by IT; comprises the data representing the legally binding, current and historical position of the business.*
- ***Machine-generated data** is the immediate output of sensors and machines that record measures and events in an increasingly instrumented and interconnected physical world—the Internet of Things, created both internally and externally to the enterprise.*
- ***Human-sourced information** comprises the messages generated directly by people to communicate their thoughts and ideas about the world, from text messages to YouTube videos, and everything in between.*

From a technical point of view, these three data domains require relatively differentiated types of storage and processing. While the actual platforms required may change as technology evolves, the distinctly different characteristics of the domains suggest that there will be three (or perhaps more) different optimal performance points on the spectrum of available technology at any time. A logical architecture based on these domains has been developed[8] and is shown in Figure 10.

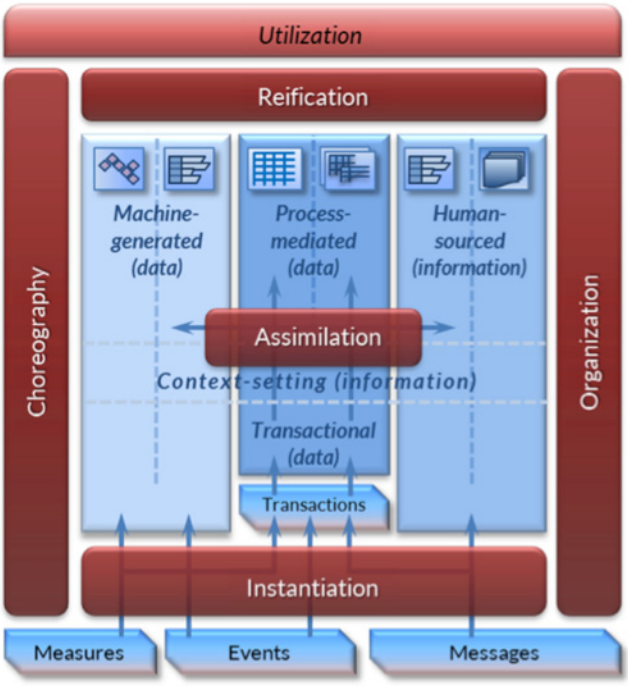


Figure 10: REAL logical architecture for Big Data.

[8] Devlin, Barry, *"Business unIntelligence: Insight and Innovation Beyond Analytics and Big Data"*, Technics Publications, New Jersey, (2013)

The Realistic, Extensible, Actionable, Labile (REAL) logical architecture describes a set of information and process components required to gather, create, manage and utilize all types of information in support of modern, real-time, information-rich businesses. In contrast to prior data warehousing, Master Data Management (MDM), operational systems and similar architectures, the REAL architecture is an enterprise architecture that explicitly includes all types of business processes and covers *all* the information used by the business.

REAL logical architecture describes a set of information and process components required to gather, create, manage and utilize all types of information in support of modern, real-time, information-rich businesses.

The three data domains appear as pillars, fed in parallel, rather than the sequentially fed layers typically shown in traditional data warehouse architectures. The central, process-oriented data pillar contains traditional operational and core informational data, fed from legally binding transactions (both OLTP and contractual). It is centrally placed because it contains core business information, including traditional operational data and informational EDW and data marts in a single logical component. **Machine-generated** and **Human-sourced** information are placed as pillars on either side. The leftmost pillar focuses on real-time and well-structured data, while the one on the right emphasizes the less structured and, at times, less timely information. In this architecture, metadata – more correctly labeled as context-setting information – is explicitly included as a part of the information resource and spans the three pillars. While these pillars are independent, their content is also coordinated and made consistent, as far as necessary, based on the core information in the **Process-mediated** pillar and the context-setting information stored in all three pillars.

Current information platform choices are:

- ***Process-mediated data:*** *General purpose relational databases, increasingly including in-memory or solid-state disks and hybrid models, that can support a mix of read-write and read-only processing.*
- ***Machine-generated data:*** *Depending on data volumes and speeds, ranging from complex event processing (streaming) systems, through NoSQL data stores, to high-performance relational database appliances.*
- ***Human-sourced information:*** *Hadoop systems, content management stores and file systems.*

In the context of Big Data, three process components play significant roles. Instantiation gathers data and information from all sources, classified here as measures, events and messages. Such gathering includes both copying data into the internal storage and process environment as well as accessing it "on the fly." Assimilation is the process that prepares and reconciles data and information prior to users accessing or using it. Assimilation corresponds to a combination of cleansing, data integration (ETL/ELT) and data virtualization tools. Reification prepares and reconciles data and information in real-time as users access and use the data; it corresponds to tools usually labeled data virtualization. Utilization includes all user tools and applications. Of most interest for Big Data today are business analytic tools.

3.4. Surfing the Sensor Data Wave

The Internet of Things[9] has been accumulating column-inches in the technical and popular media over the past few years. Much of the focus has been on the sensors and machines that record and transmit data about an ever-growing array of measures and events in the real world. In the automobile industry, for example, lower-cost automobile models already incorporate dozens of sensors for engine temperatures, oil pressures, timings, torques and more. Luxury models have hundreds of sensors that go far beyond the operation of the vehicle. Energy utilities are incorporating sensors throughout their supply networks,

[9] http://en.wikipedia.org/wiki/Internet_of_things

from production facilities all the way to domestic meters, all calling home, often wirelessly, with interesting data points. From simple RFID[10] tags to smartphones that carry GPS[11] location sensors, accelerometers and more, organizations are creating an enormous infrastructure that can capture a limitless number of physical characteristics. Much of the hype has focused simplistically on the growth in device numbers and penetration. Simple applications automate and accelerate existing measurement processes; smart meters are an obvious example. Other applications allow businesses to reinvent their processes, as can be seen with usage-based automobile insurance.[12] What has not been clear so far is how far this data and these applications have penetrated into everyday business computing. The popular press and many hardware and software vendors predict that the Internet of Things will become the dominant source (by volume) of Big Data. But when?

Many predict that the Internet of Things will become the dominant source (by volume) of Big Data. But when?

According to the EMA/9sight respondents, the answer is that it has already – if not yet by actual volume, then certainly in terms of percentages of types of data included in Big Data projects. We classified data types according to the above-described, tri-domain model. The Internet of Things creates **Machine-generated** data according to this classification. The class also includes internally generated sensor and machine-to-machine information: raw data ending up in web logs, telecommunications CDR stores and so on. The other major external source of Big Data is social media and similar information, called **Human-sourced** information.

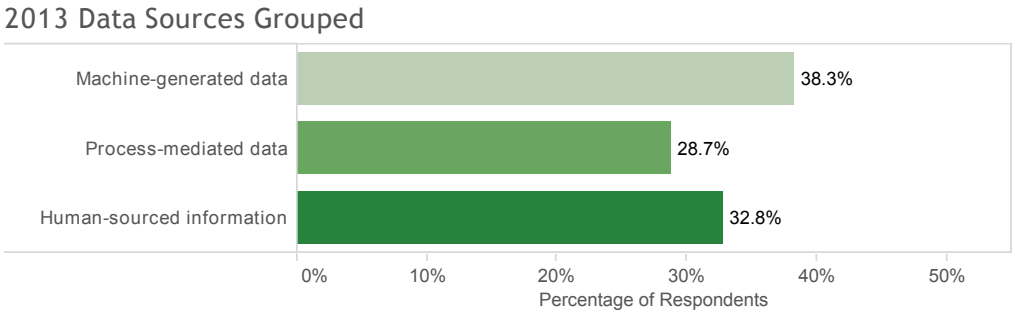


Figure 11

As can be seen in Figure 11, the overall percentages have switched from just under 24% to over 38% for **Machine-generated** data sources. **Human-sourced** information made a similar change. In 2012, **Human-sourced** information was indicated by over 45% of the respondents and moved to less than 33% in 2013.

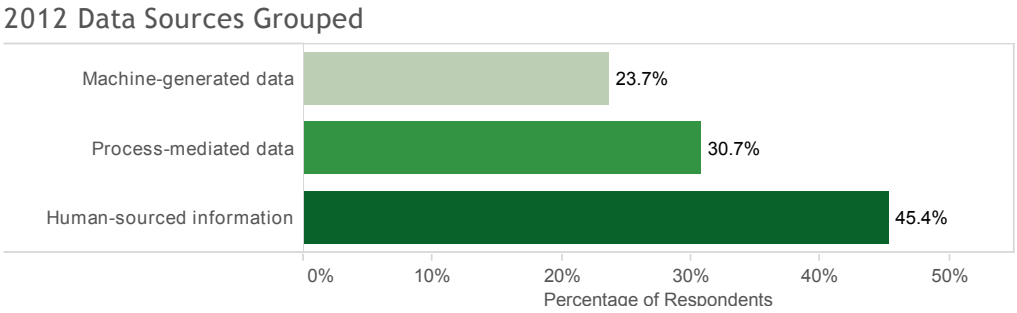


Figure 12

[10] http://en.wikipedia.org/wiki/RFID
[11] http://en.wikipedia.org/wiki/GPS
[12] http://en.wikipedia.org/wiki/Usage-based_insurance

While the response set did not distinguish between internally and externally sourced **Machine-generated** data, EMA/9sight believes internal sources are unlikely to have grown substantially in the intervening year.

3.5. What is Next for Big Data? Ethics!

From a pure technology viewpoint, it seems reasonably certain that current trends will continue. Emerging Big Data toolsets will continue to evolve towards supporting real-time use cases. These toolsets will also expand support for Big Analytics. The former is simply following the same trajectory seen in the earliest days of computing when online computing replaced batch for most needs. EMA/9sight expects to see this drive a gradual reduction in the plethora of competing, and often overlapping, utility functions and add-ons in the Hadoop ecosystem. Similar demands for real-time analytics will likely reinforce this trend.

While there is a slowly growing emphasis on data management, this is unlikely to deliver significant support in the short- to medium-term. Both business and IT stakeholders will continue to be faced with a smorgasbord of poorly integrated tools and "Do It Yourself" (DIY) data management environments. The best hope for those who value data quality and consistency will still reside in the more traditional relational and content management environments. The relational data store vendors will continue to integrate Big Data technology into the traditional environment in multiple ways. This will include front-ending Hadoop with relational tools and incorporating support for NoSQL stores into the relational databases.

Another issue for Big Data in the coming couple of years will not be technical in nature. It will be ethical. Every industry is rushing to profile customers – in some cases, in the interests of offering better service or personal convenience. However, the reality is less comforting. Profiling enables ever more subtle customer segmentation, which can easily turn to discrimination in pursuit of profit. Gathering and combining Big Data from every conceivable source may enable the type of discriminatory practices based on race, religion, medical condition or sexual preference that have been illegal in the physical world (depending on the jurisdiction) for years. Combined with personalization of browsing, the effect is so subtle and endemic that those being excluded from particular programs or benefits may not even know that the programs exist.

Another issue for Big Data in the coming couple of years will not be technical in nature. It will be ethical.

EMA/9sight hope to see new approaches to protecting privacy and controls on how data is combined and used can avert the worst of such anti-democratic practices. However, the responsibility rests squarely on the shoulders of the organizations that are defining and implementing these Big Data programs. Or perhaps drive a renewed focus on using Big Data for the scientific purposes from which the term originated, and solving some of the pressing survival issues we face in today's world.

4. Who's Who of Big Data

The 2013 EMA/9sight survey respondents were selected from a wide range of industries, company sizes and geographic distribution. This diversity provided a well-balanced look at the makeup of data management technologists and business stakeholders embracing Big Data around the world.

4.1. Enterprising Company Size

The EMA/9sight survey examined companies across a continuum of size. Corporate headcount is distributed in the following manner:

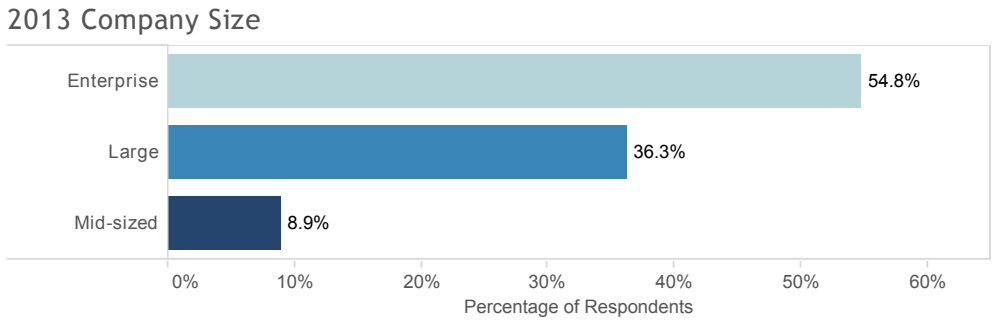


Figure 13

This distribution is significantly different from the distribution of company size from the 2012 survey respondents displayed in Figure 14:

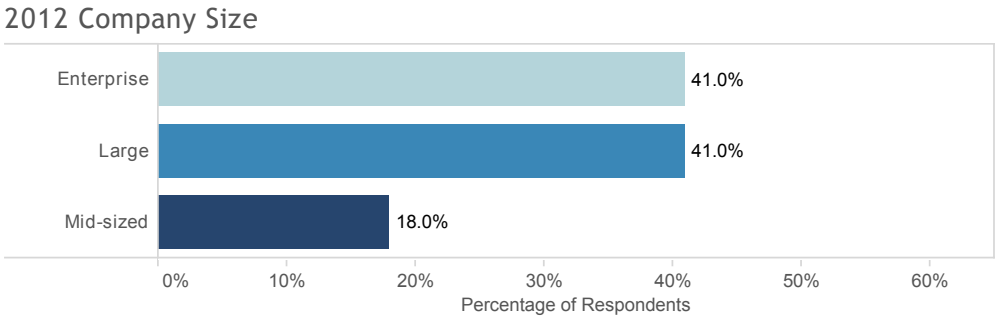


Figure 14

The increase in **Enterprise** companies (5000+ headcount) and the decrease in **Mid-sized** (less than 500 headcount) organizations can be attributed to many factors. One of these factors may be a lower response rate from the **Media & PR** industry segment (see Industry Segment definitions below). Many of the 2012 **Mid-sized** respondents came from this industry segment.

4.2. Industrial Strength

Any investigation of Big Data initiatives should take into consideration the various industries and industry segments associated with the respondents. Some industries are on the cutting edge of developments while others are still gaining traction with their Big Data initiatives.

In this year's study and in the 2012 EMA/9sight Big Data research, industries were grouped into the following designations:

- ***Finance:*** *Finance, Banking, and Insurance*
- ***Public Services:*** *Government, Education, Non-Profit/Not for Profit, and Legal*
- ***Manufacturing:*** *All non-Computer or Networking related Manufacturing industries*
- ***Industrial:*** *Aerospace and Defense manufacturing, Oil and Gas production and refining, Chemical manufacturing, and Transportation and logistics organizations like Airlines, Trucking and Rail*
- ***Leisure:*** *Hospitality, Gaming and Entertainment, as well as Recreation and Travel*

- ***Media and PR:*** *Marketing, Advertising, Public Relations and Market Research, and Publishing and Broadcasting*
- ***Utilities Infrastructure:*** *Telecommunications Service Providers, Application, Internet and Managed-Network Service Providers, and Energy production and distribution Utilities*
- ***Retail:*** *End Consumer Retail and Wholesale and Distribution*
- ***Healthcare:*** *Medical device and supply and Pharmaceutical production*

In 2013, the number of **Leisure** and **Media & PR** industry segments respondents was not statistically significant among the respondents. Therefore, these industry segments are not included in the industry-based analysis for 2013. However, you will see those segments represented in 2012 result sets.

With this in mind, the breakdown by industry segment for the 2013 panel respondents is as follows:

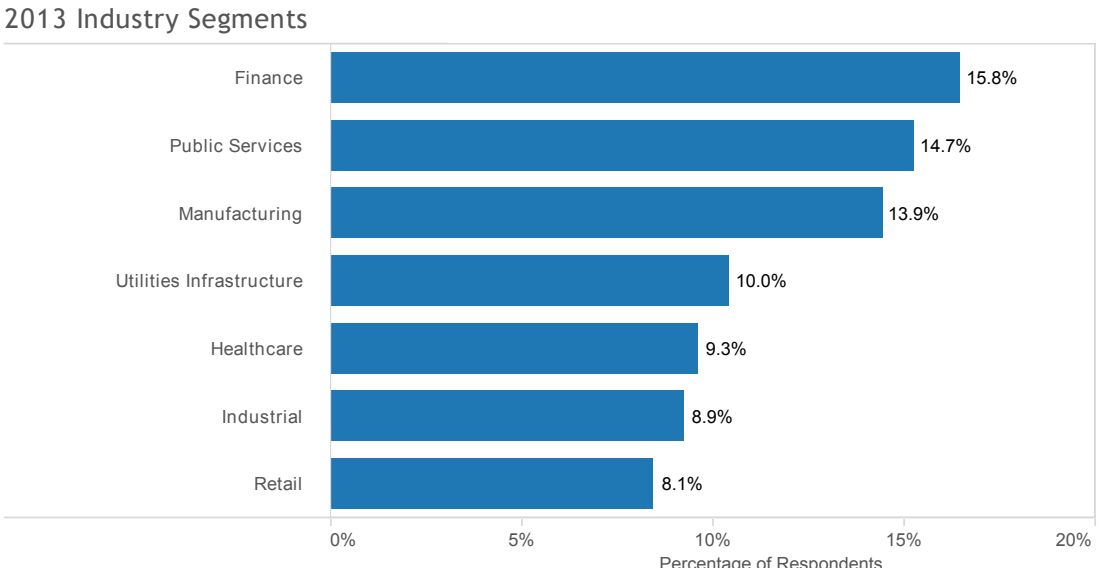


Figure 15

The survey participation of the **Manufacturing** industry segment increased significantly between 2012 and 2013.

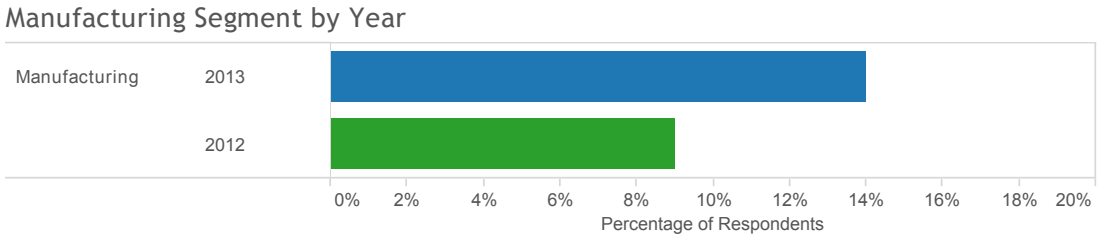


Figure 16

This increase demonstrates the maturing nature of Big Data, in particular in the area of sensor based data collection, as part of machine-to-machine communications.

The size of the companies comprises the following percentages based on industry segments:

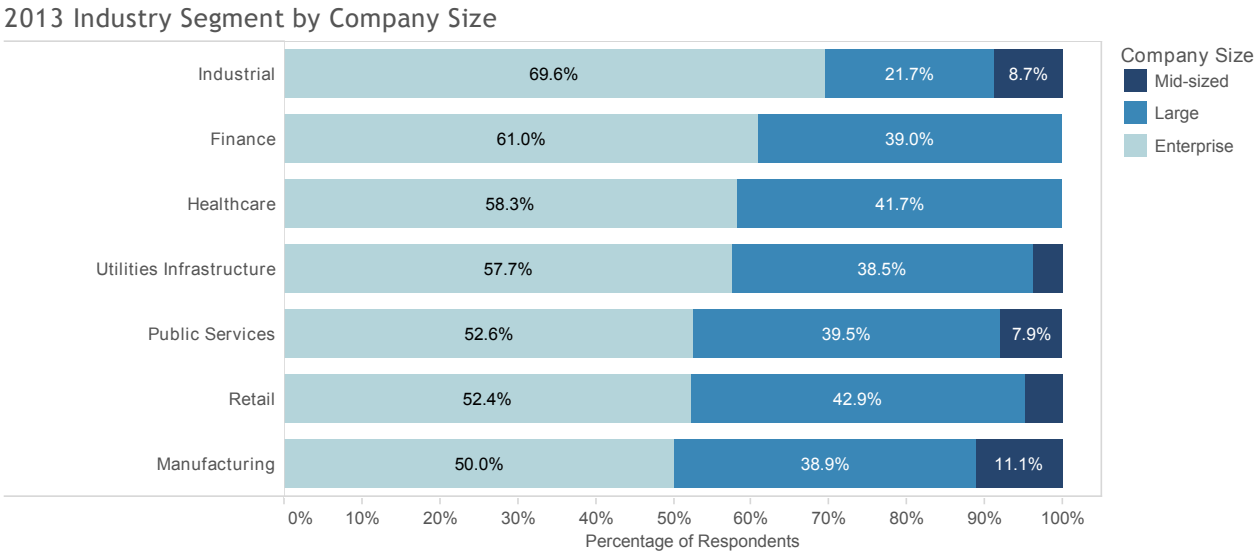


Figure 17

Not surprisingly, **Enterprises** dominate the **Industrial**, **Finance** and **Healthcare** industry segments. It is interesting to see **Mid-sized** organizations with relatively large percentages of **Manufacturing** and **Industrial** segments working with Big Data initiatives.

4.3. Around the Globe

The comparison between the 2012 and 2013 geographic distributions demonstrates that Big Data is not a geographically isolated trend to one part of the globe.

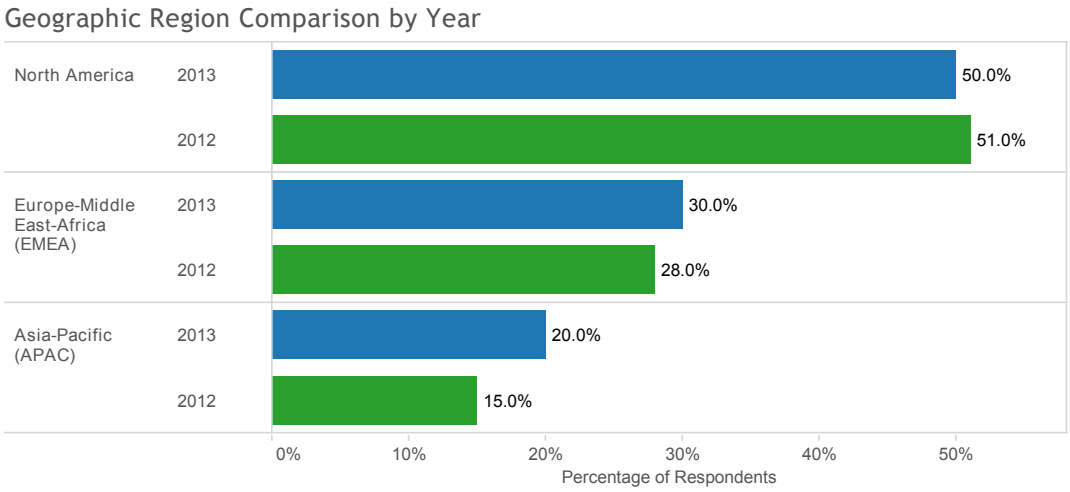


Figure 18

Between 2012 and 2013, **North American** and European-Middle East-Africa (**EMEA**) region respondents are approximately the same in distribution. There are significantly more Asia-Pacific (**APAC**) region respondents in 2013 than in 2012. This change in distribution may be due to the fact that there was a significant decrease in the number of Central and Latin American (**CALA**) region respondents between 2012 and 2013.

4.4 Corporate Innovation

As part of the 2013 survey, EMA/9sight asked respondents to describe their corporate cultures in terms of the standard Rogers Adoption Curve[13]. The Rogers Curve includes the following segments:

- ***Innovator:*** *Brave. Initiating change.*
- ***Early adopter:*** *Try out new ideas but in a careful way.*
- ***Early majority:*** *Thoughtful. Accepting change more quickly than average.*
- ***Late majority:*** *Skeptical. Uses new ideas only when majority are using it.*
- ***Laggard:*** *Critical towards new ideas. Only accept change when it has become mainstream.*

As can be expected, the distribution of survey respondents matches the normal distribution of a Roger Curve with a majority of respondents in the Early Majority segment and less respondents toward either end of the scale.

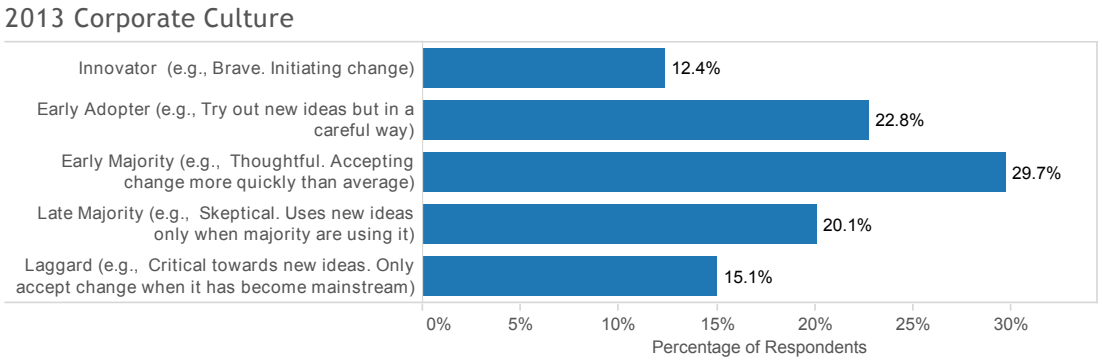


Figure 19

The distribution of respondents along with their geographic region shows that a high percentage of respondents from the **EMEA** and **APAC** regions consider themselves to have a corporate culture in the **Innovation** stage. This distribution is higher than their overall geographic distribution for this research.

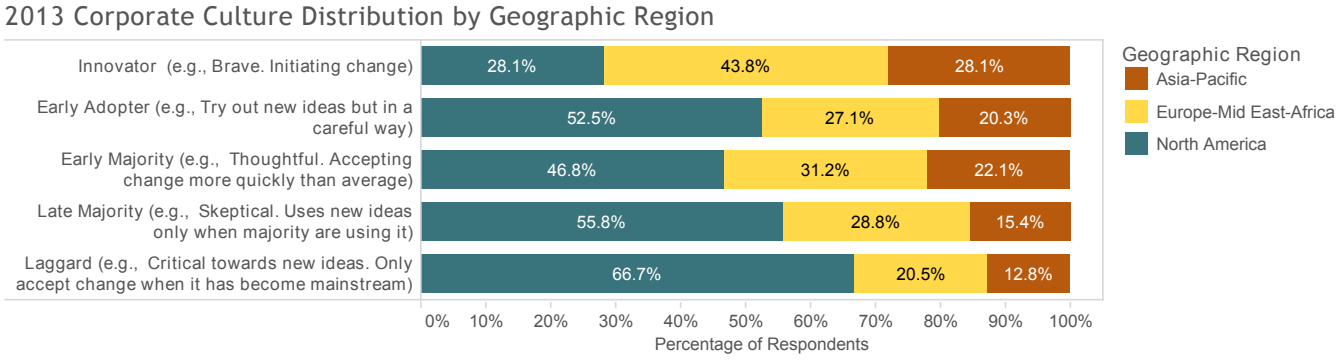


Figure 20

As you approach the more mature segments of the Rogers Adoption curve, the distribution returns expected bands associated with the overall geographic distribution. Surprisingly, there are a relatively high percentage of **North American** respondents who consider themselves to be in the **Laggard** stage. This is interesting because it is widely held that the Big Data "revolution" was started in North America. This shows the disparate nature of the perception of Big Data adoption and innovation not just around the world, but also within a particular region.

[13] http://en.wikipedia.org/wiki/Technology_adoption_lifecycle

4.5. The Case for Big Data

EMA/9sight asked respondents how they are using or planning to use their Big Data implementations, offering a selection of **Use Case** options. Respondents were given the opportunity to select multiple **Use Cases** as appropriate to their Big Data initiative, as they might be engaged in multiple paths as part of their Big Data program.

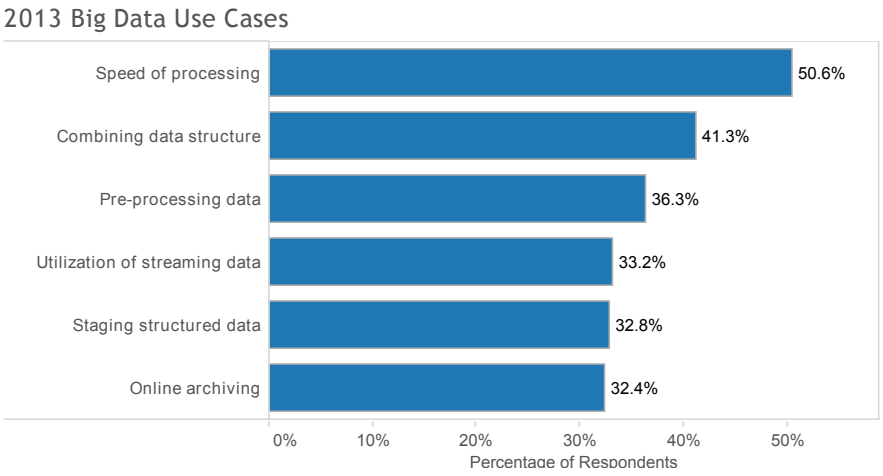


Figure 21

Over 50% of respondents indicated that **Speed of Processing Response** was among their Big Data initiative **Use Cases**. This is a significant increase from 2012.

Comparison of the common responses between the 2012 and 2013 surveys shows the most common response from 2012 – **Online Archiving** – has fallen dramatically and has been replaced almost as significantly by **Speed of Processing Response**.

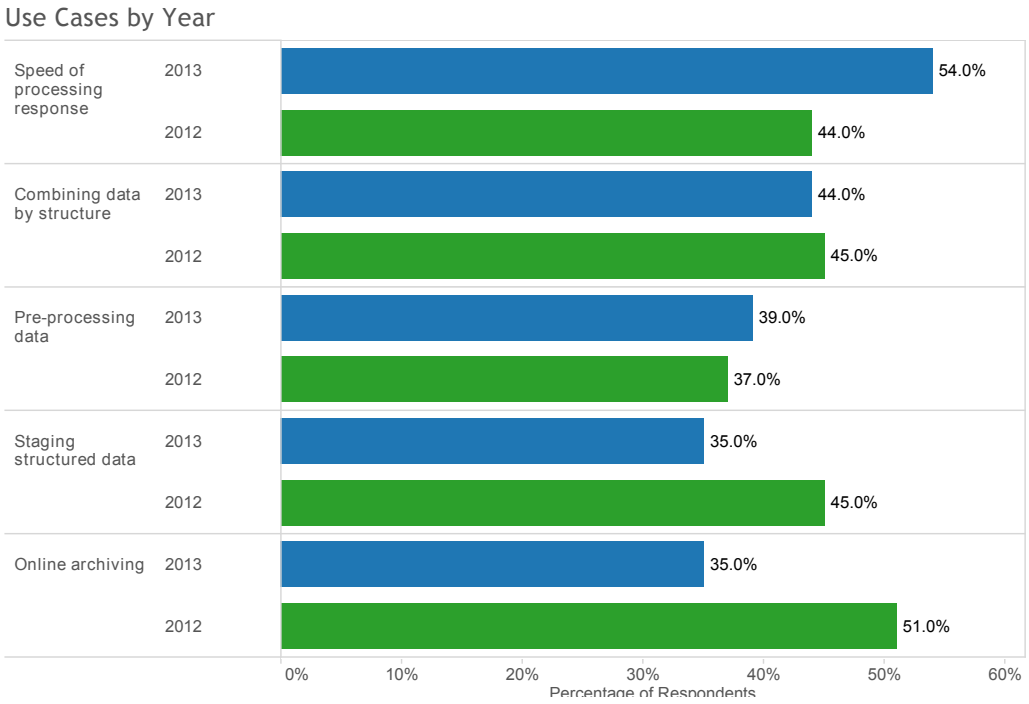


Figure 22

This change from 2012 to 2013 shows the maturing of the respondents' Big Data initiatives. Based on 2012's research, EMA/9sight concluded that the **Online Archiving** use case was an early stage of a Big Data initiative. Building skills and establishing the data management practices necessary to engage with new data formats and processing complexities were critical in 2012. In 2013, those initial steps have been addressed and companies are starting to understand how Big Data initiatives can influence their organizations. The skill gap, however, continues to be a challenge for practitioners of Big Data project. The ascension of **Speed of Processing Response** indicates that the research respondents are looking to gain competitive advantage via workloads that stress fast reaction time and low latency.

4.5.1. Breaking Down Industry Cases

When Use Cases by industry segment are broken out, the respondents in **Finance** and **Retail** indicated the highest interest in Speed of Processing Response for their use cases. This observation goes along with the nature of those industries and the real-time nature of their operational business models.

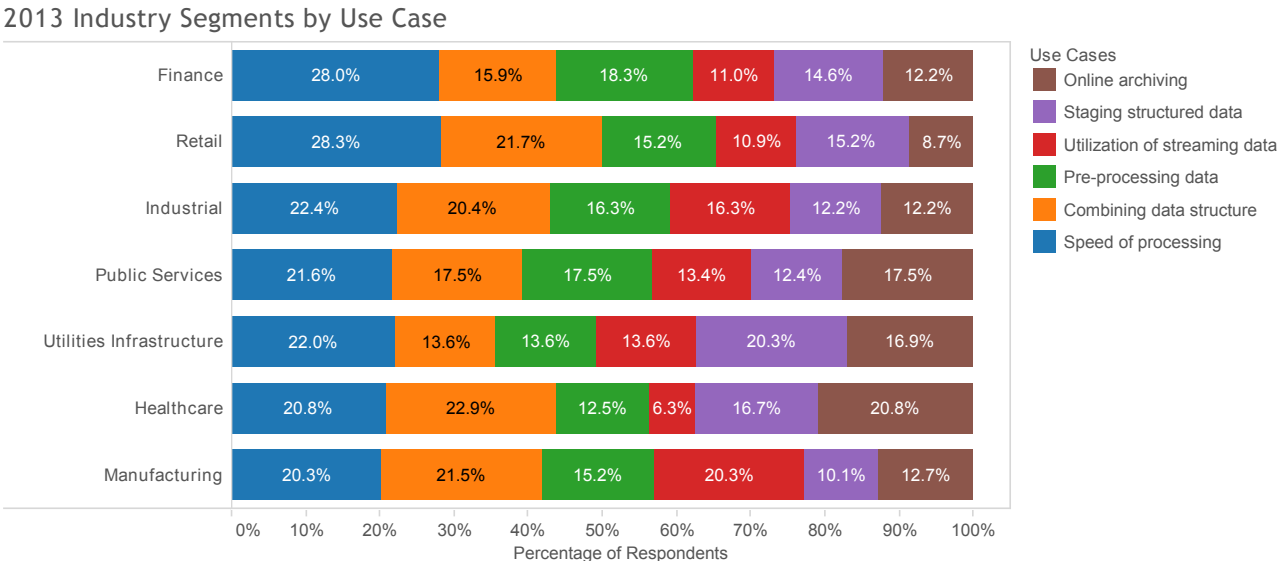


Figure 23

The **Healthcare** industry segment showed an interesting set of responses favoring **Combining Data by Structure** as well as **Speed of Process Response**. As **Healthcare** moves into an era of applying greater levels of technology and needing to bridge to the gap between **Process-mediated** data in operational systems and **Human-sourced** information from medical charts and forms, this will be an opportunity to watch for in organizations across the field of **Healthcare**.

4.6. Managing Hurdles

With any new set of technologies or initiatives, there are a considerable number of issues that need to be addressed and/or overcome for a successful implementation. EMA/9sight panel respondents were asked to identify the obstacles to their Big Data initiatives, cultural considerations of **Stakeholder Issues** and **Strategy Issues** were the top two responses.

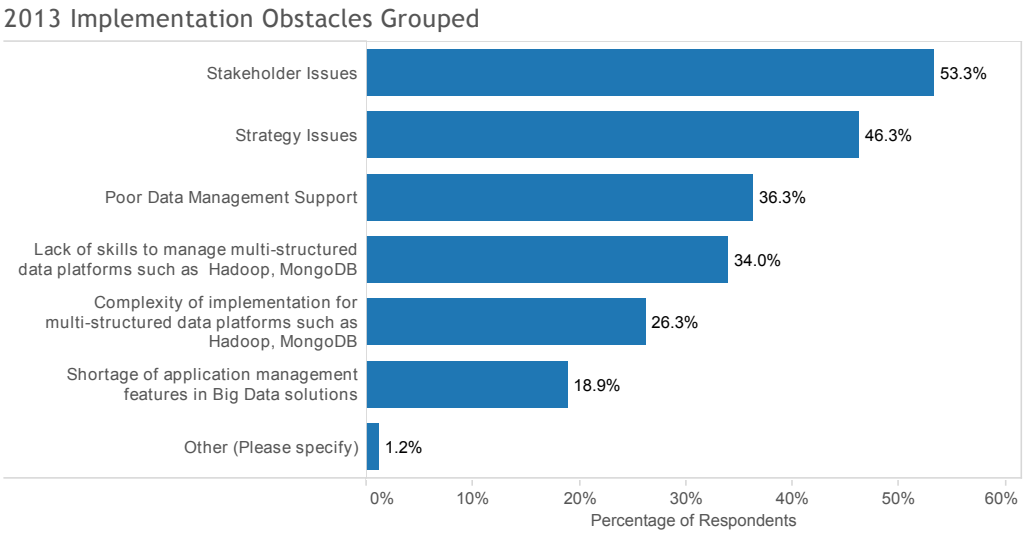


Figure 24

Stakeholder Issues of communicating with management and building the case for their approval or buy-in, and **Strategy Issues** relating to aligning on implementation strategy are not exclusive to the world of Big Data initiatives. These concepts are pervasive throughout business, and in particular, IT projects.

Looking into the industry segment breakdown, **Healthcare** and **Utilities Infrastructure** led in the area of **Stakeholder Issues** of communication and buy-in.

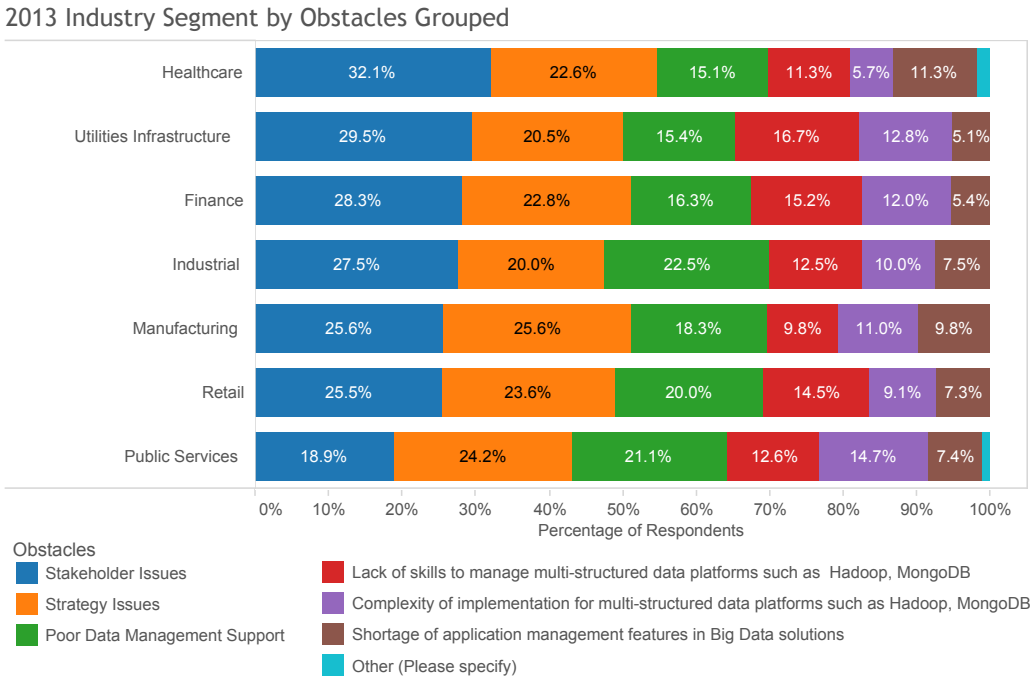


Figure 25

The **Manufacturing** and **Public Services** industry segment respondents had significant obstacles relating to **Strategy Issues** of alignment between stakeholders, whether they are associated with business or technology, once buy-in had been achieved.

5. Of Projects and Programs

In the 2012 EMA/9sight Big Data study, EMA/9sight asked respondents about the implementation stage of their Big Data initiative. Since the survey asked about their project(s) taken together, EMA/9sight received indications that showed that there was actually an ongoing program of projects at various stages of implementation. For 2013, EMA/9sight made the decision to delve deeper into the status and characteristics of those individual projects.

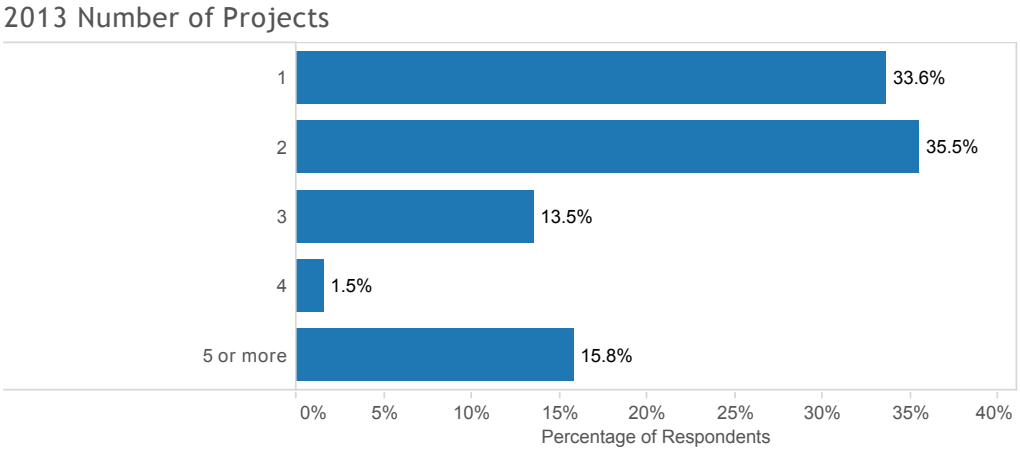


Figure 26

This year's research collected data on 597 active, or ongoing, Big Data projects. These projects are implemented on a variety of platforms that support both analytical and operational workloads. EMA/9sight analysis illustrates a clear shift in the Big Data space toward maturing adoption, greater sophistication of workloads and strategies, along with a significant move toward projects that are in the **In Operation** stage. Of the companies surveyed, 67% indicated they were working on two or more projects with 16% of the companies working on five or more.

This year's research collected data on 597 active, or ongoing, Big Data projects.

5.1. Developing Maturity

Project stages are classified into three phases: **In Operation**, **Serious Planning** and **Investigating**. Each of these identifies a different phase of implementation for an active Big Data project. Of the projects in this year's research, 53% are **In Operation**, over double the number of projects in the **Serious Planning** or **Investigating** categories.

The following groupings are used to define an organization's status in relation to its Big Data initiatives:

- ***In Operation:*** *This represents actual implementations of Big Data projects including "Already having a project in production" and "Currently working to implement a pilot project." These respondents have hands-on experience with both Big Data business requirements and technologies that solve those requirements.*
- ***Serious Planning:*** *This represents near-term to immediate Big Data projects. These include survey respondents who indicated planning for implementation within one to six months. This group represents organizations that are close to or on the verge of signing contracts for hardware and software licenses associated with their Big Data implementation.*
- ***Investigating:*** *This grouping represents those organizations still looking at Big Data requirements and Big Data technologies. These respondents are 7+months out from implementing a Big Data solution.*

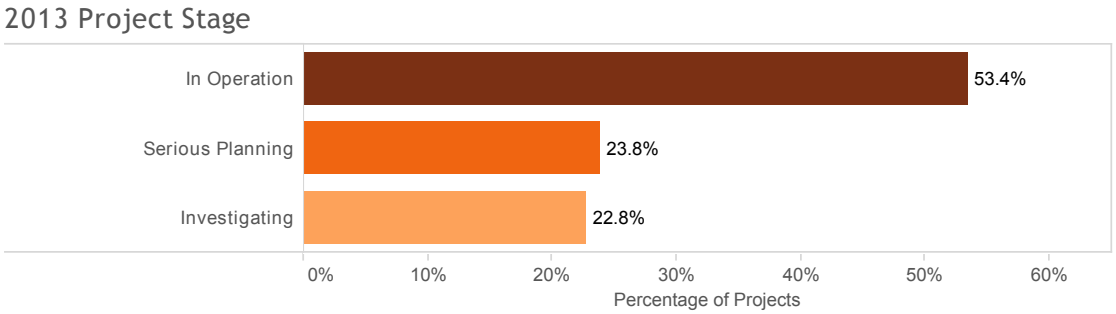


Figure 27

A significant shift in project stages can be seen from 2012 – 2013. It is clear that the 2013 respondents are further along in their Big Data efforts than in 2012. Respondents indicated 17% more of their initiatives are **In Operation** over the previous year. In comparison, **Serious Planning** and **Investigating** are both lower than last year's indications. The increase in **In Operation** projects points to a growth in adoption and a maturing of the market.

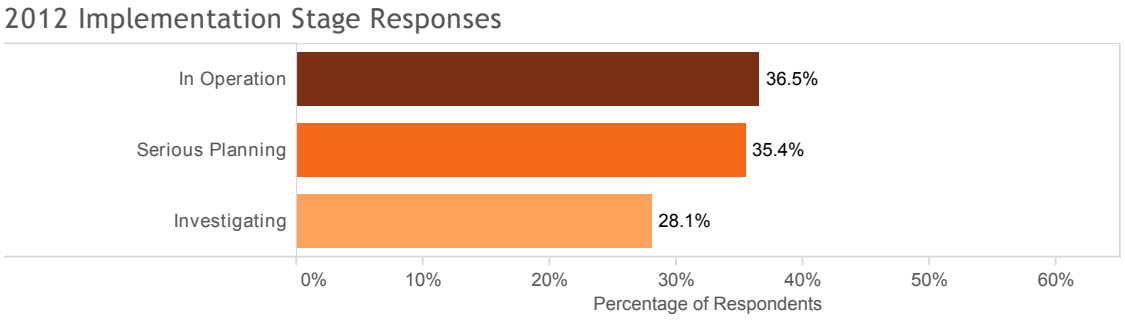


Figure 28

5.1.1. Implementing Projects

There are identifiable trends in how company size influences Big Data projects. This research categorizes company size into three groups, **Mid-sized**, **Large** and **Enterprise** organizations. In 2012, the results were somewhat flat and did not identify dramatic trends as to how company size influenced where a company was with regard to project stage.

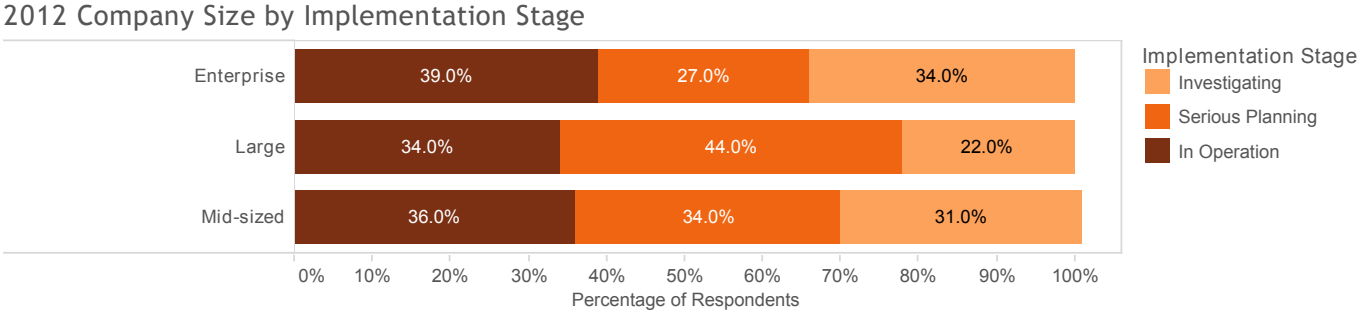


Figure 29

In 2013, the data shifted dramatically and showed that **Enterprise** firms are delivering more Big Data projects. At **Enterprise** sized companies, 56% of the active projects are **In Operation**.

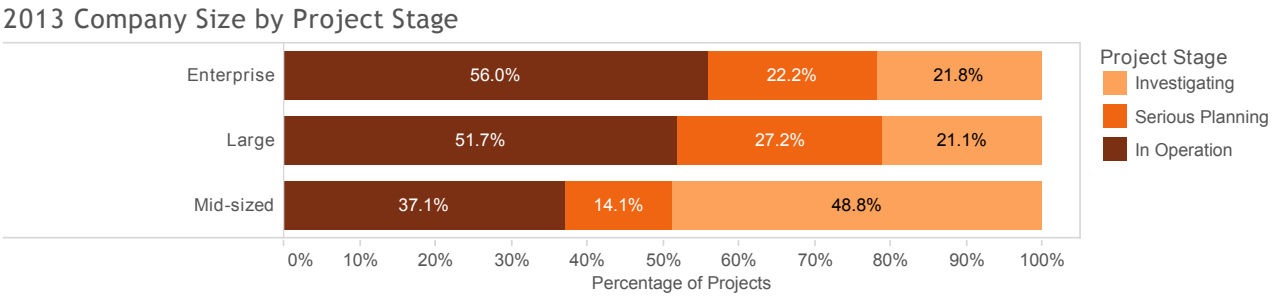


Figure 30

Large sized company projects are moving quickly as well, showing an 18% jump from 2012 with **In Operation** efforts. Conversely, **Mid-sized** companies are moving more slowly with their projects. Respondents from **Mid-sized** organizations indicated an increase in efforts in the **Investigating** stage and are seven or more months away from completion.

Enterprise sized companies have the most active Big Data programs. **Enterprises** are more likely to be working on multiple Big Data projects than **Large** and **Mid-sized** companies put together.

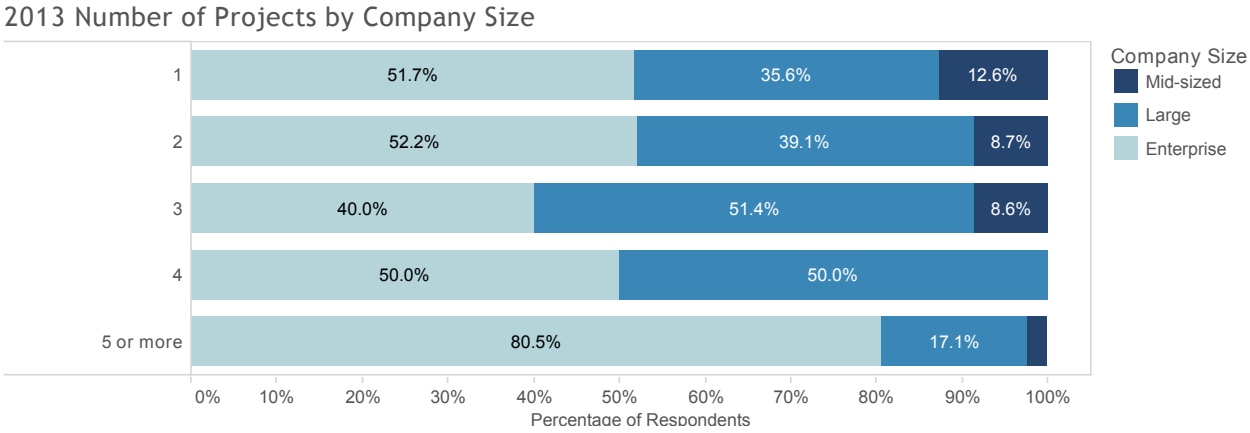


Figure 31

Large and **Mid-sized** firms are not nearly as aggressive in their pursuit of larger Big Data programs of five or more projects.

5.1.2. Looking Across Industries

Most industry sectors have identified opportunities to deploy Big Data projects. In 2012, EMA/9sight survey respondents from **Media & PR**, **Retail**, **Industrial** and **Finance** industry segments were most likely to have some aspect of their Big Data initiative **In Operation**. In 2013, **Healthcare** leads all industry sectors for projects **In Operation** followed by **Retail**, **Industrial** and **Manufacturing**. The **Public Services** industry segment seems to be the most cautious with 39% of its projects still in the **Investigating** category. This may point to new interest in this sector for Big Data projects. The **Finance** industry placed lower than expected for **In Operation** projects, perhaps pointing to more complicated and difficult-to-deploy projects. In the EMA/9sight 2012 research, this industry segment was identified as an early adopter of Big Data technology, and the slowdown could be a representation of a natural pause in innovation as it reaps the rewards of previous projects.

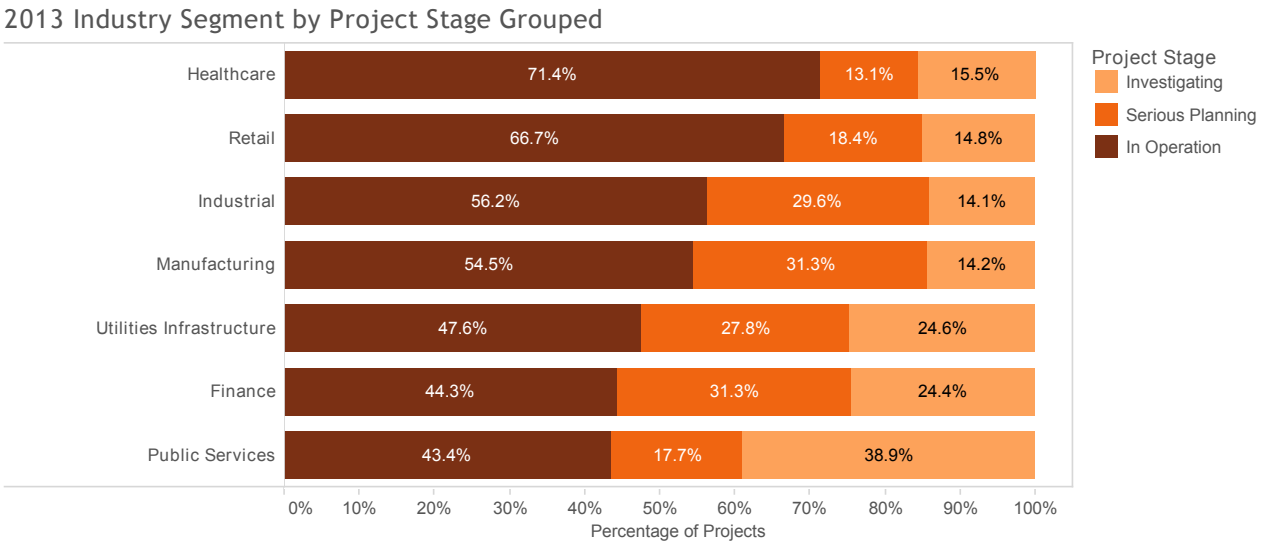


Figure 32

In 2012, EMA/9sight respondents came from a variety of industry sectors. **Leisure**, **Media & PR**, **Utilities Infrastructure** and **Manufacturing** all indicated they were predominately in the **Serious Planning** stage indicating these industries should have moved forward to **In Operation** status this year. EMA/9sight research in 2013 switched from respondents within industries to projects in industries, resulting in a clearer view of industry maturation and adoption direction.

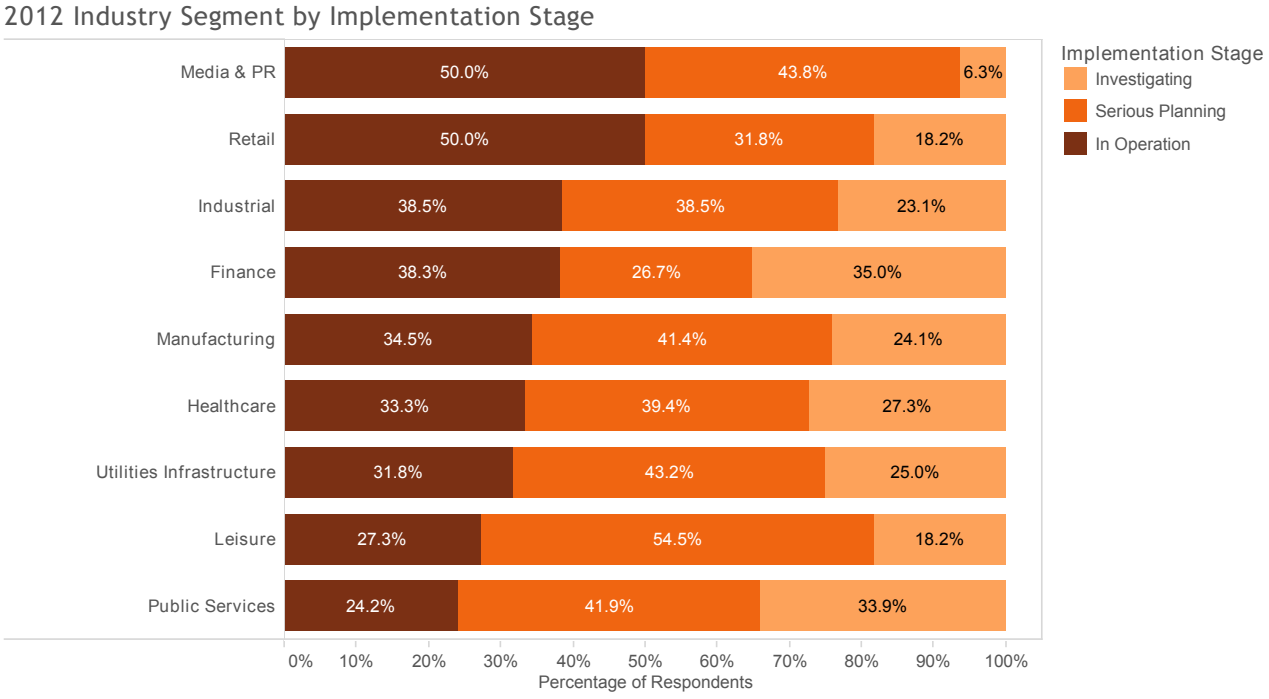


Figure 33

5.2. Meeting the Challenge

As more initiatives are implemented, these companies are working to deliver critical and sophisticated projects to their internal and external stakeholders. Most of these workloads incorporate multiple data sources, multi-structured as well as structured in nature.

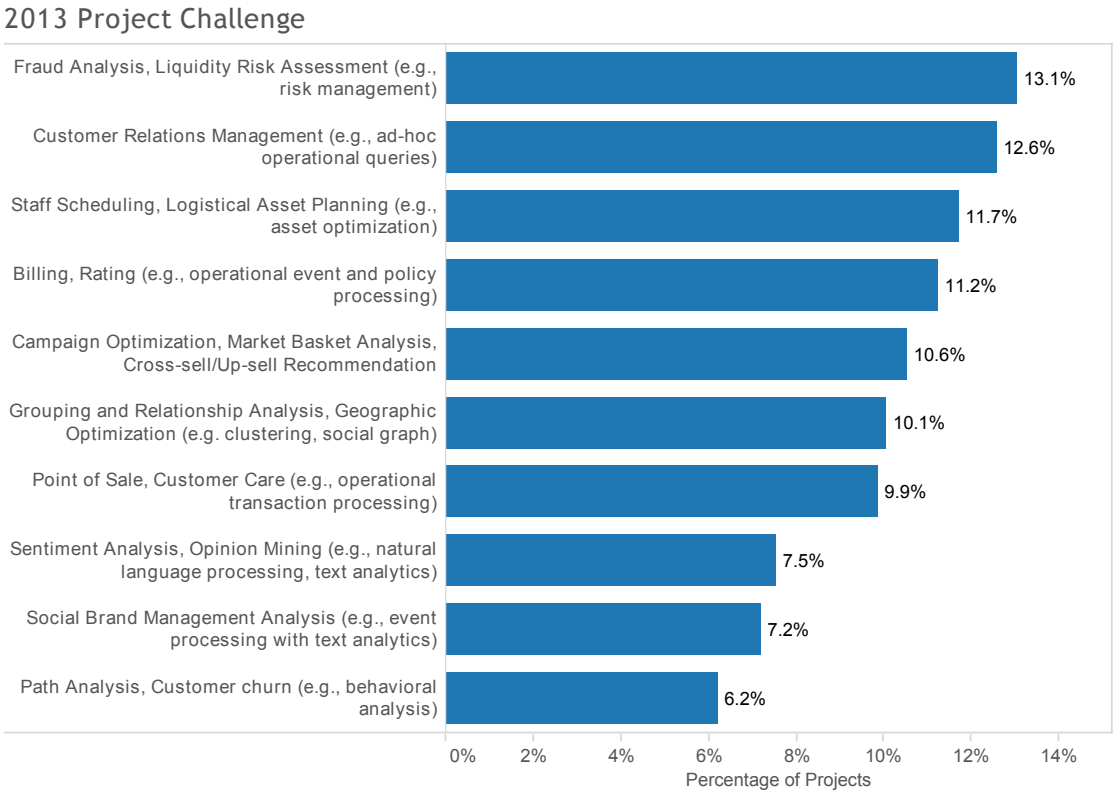


Figure 34

The project characteristics above show that many of these projects rely on real-time insights and processes. **Fraud Analysis, Liquidity Risk Assessment (e.g., risk management)**, **Staff Scheduling, and Logistical Asset Planning (e.g., asset optimization)** project challenges, as they are integrated into operational processes, require a lower latency of response. This indicates that Hadoop, with its current batch oriented processing model via MapReduce, is not the only platform being utilized.

5.2.1. Putting Them Together

As the EMA/9sight survey examined the nearly 600 active projects, groupings of challenges were quickly identified:

- ***Operation Analytics:*** *This includes the growing area where operational processes are merged with analytical results to improve overall business performance with the following types of workloads: Campaign Optimization, Market Basket Analysis, Cross-sell/Up-sell Recommendation; Customer Relations Management (e.g., ad-hoc operational queries); Staff Scheduling, Logistical Asset Planning (e.g., asset optimization); Fraud Analysis, Liquidity Risk Assessment (e.g., risk management).*
- ***Operational Processing:*** *Big Data initiatives include the beating heart of most organizations. The operational processes associated with Billing, Rating (e.g., operational event and policy processing) Point of Sale, Customer Care (e.g., operational transaction processing).*

- ***Relationship Analysis:*** *One of the key areas of Big Data projects is the ability to take new data sources and include them into the task workloads. Grouping and Relationship Analysis, Geographic Optimization (e.g. clustering, social graph) and Path Analysis, Customer Churn (e.g., behavioral analysis) are examples of these business challenges.*
- ***Social Brand and Sentiment Analysis:*** *One of the core functions of Big Data has been the ability to ingest and process information coming from social and sentiment data sources associated with social media and* ***Human-sourced*** *information. Sentiment Analysis, Opinion Mining (e.g., natural language processing, text analytics) and Social Brand Management Analysis (e.g., event processing with text analytics) are in this category.*

Operational Analytics leads all other categories with 47.9% of respondents stating they were executing projects around this workload.

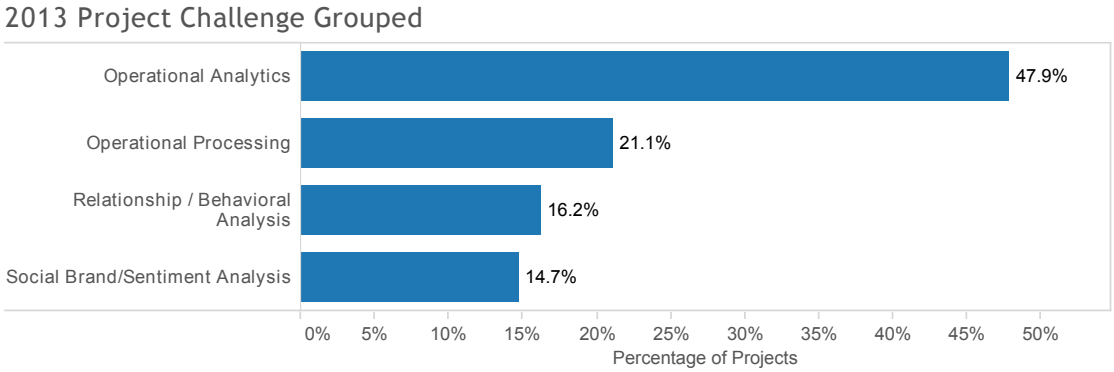


Figure 35

Diving deeper into these challenges, it is clear to see that the **Industrial** segment leads all others with the need to deliver **Operational Analytics**. This industry is often challenged by legacy systems and massive amounts of **Machine-generated** data that are difficult to manage and even more difficult to integrate into operational process workflows.

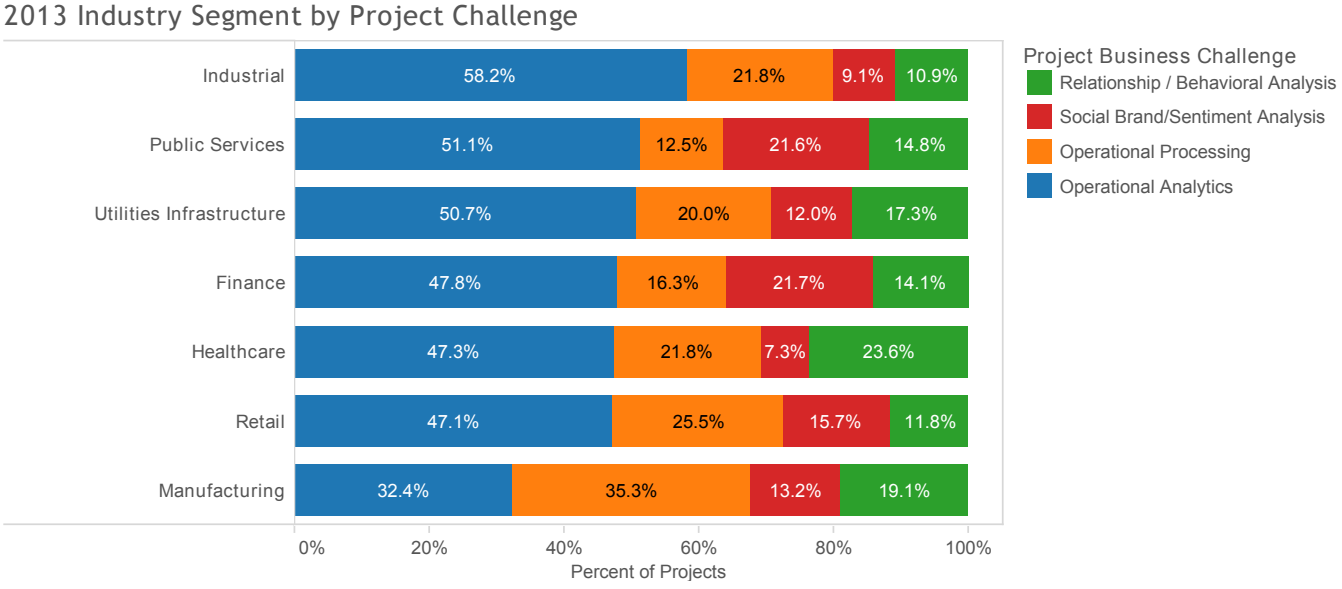


Figure 36

The **Public Services** industry segment sector is second on this list at 51%, and it too faces the challenge to integrate new data sources into its public sector related workflows. A close third in industry segments utilizing **Operational Analytics** is the **Utilities Infrastructure** segment. These organizations are moving toward more responsive water, power and network grids. The key to these initiatives is making asset and network optimization part of everyday processes.

5.3. Information Consumers

Understanding the consumer end of Big Data projects is important. These project initiatives are often complex and expensive. It makes sense that they are part of critical business processes. To that end, **Line of Business Executives** are the largest user/consumer of these projects. **Marketing and Financial Analysts** are second in this category, and often sophisticated use cases that drive Big Data initiatives.

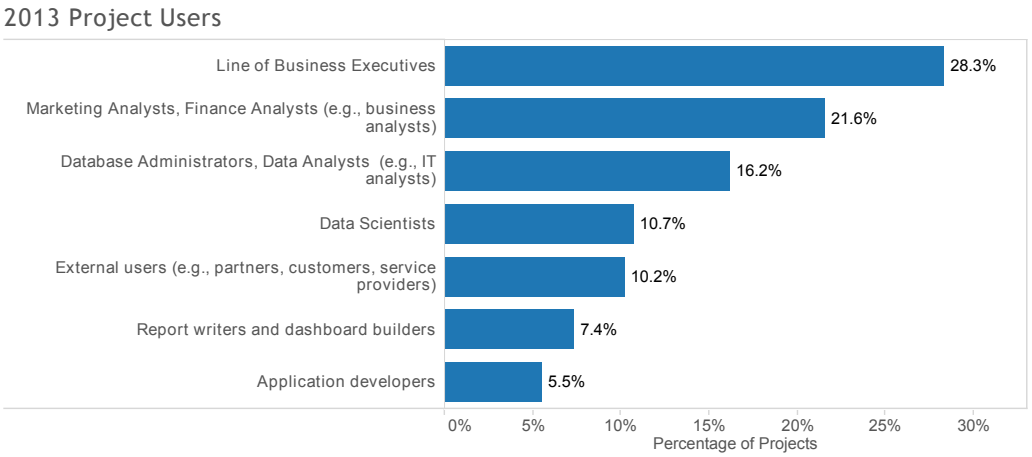


Figure 37

The highly promoted and often discussed **Data Scientist** is involved in "only" 11% of the projects examined for this research. The role of **Data Scientist** is often extremely specialized and difficult to fill due to a wide variety of necessary skill sets. **Enterprise** companies are more likely than **Mid-sized** firms to employ a **Data Scientist**. Surprisingly, though, **Large** companies have the highest percentage of **Data Scientists** as project users.

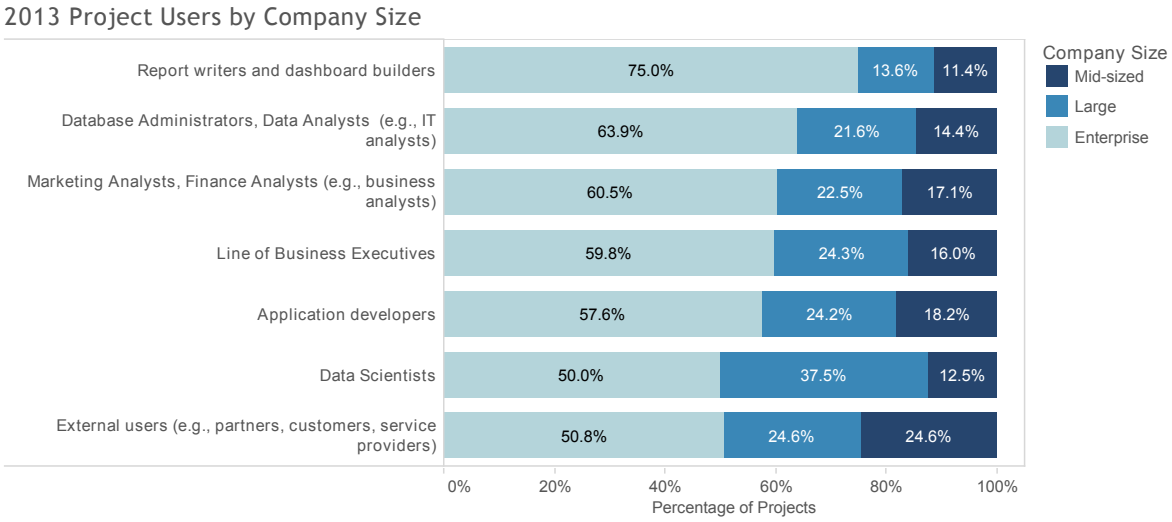


Figure 38

When comparing these two charts (2013 in Figure 38 and 2012 in Figure 39), take into consideration that the 2013 results represent project users of the 597 active projects while the 2012 results identify the aggregate user roles associated with Big Data initiatives of last year's research.

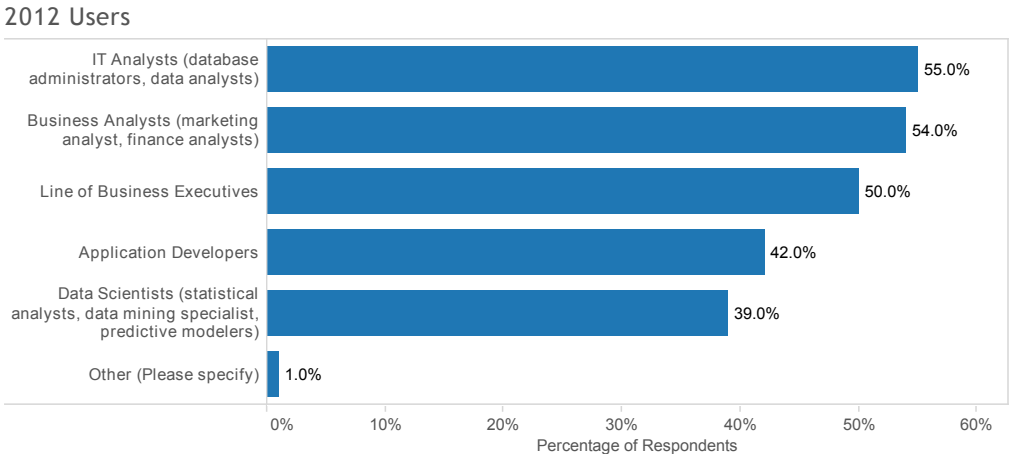


Figure 39

While the two data sets are somewhat different, the growth in **Line of Business Executives** points to a shift toward users/consumers and is supported by deeper and more significant projects.

5.3.1. Different Industries, Different Users

EMA/9sight drilled deeper into the information users associated with Big Data projects by looking at the distribution by industry segment. Several highlights include the significant lead by **Line of Business Executives** in the **Healthcare** segment. This eclipses each of other **Healthcare** user groups by at least 25%.

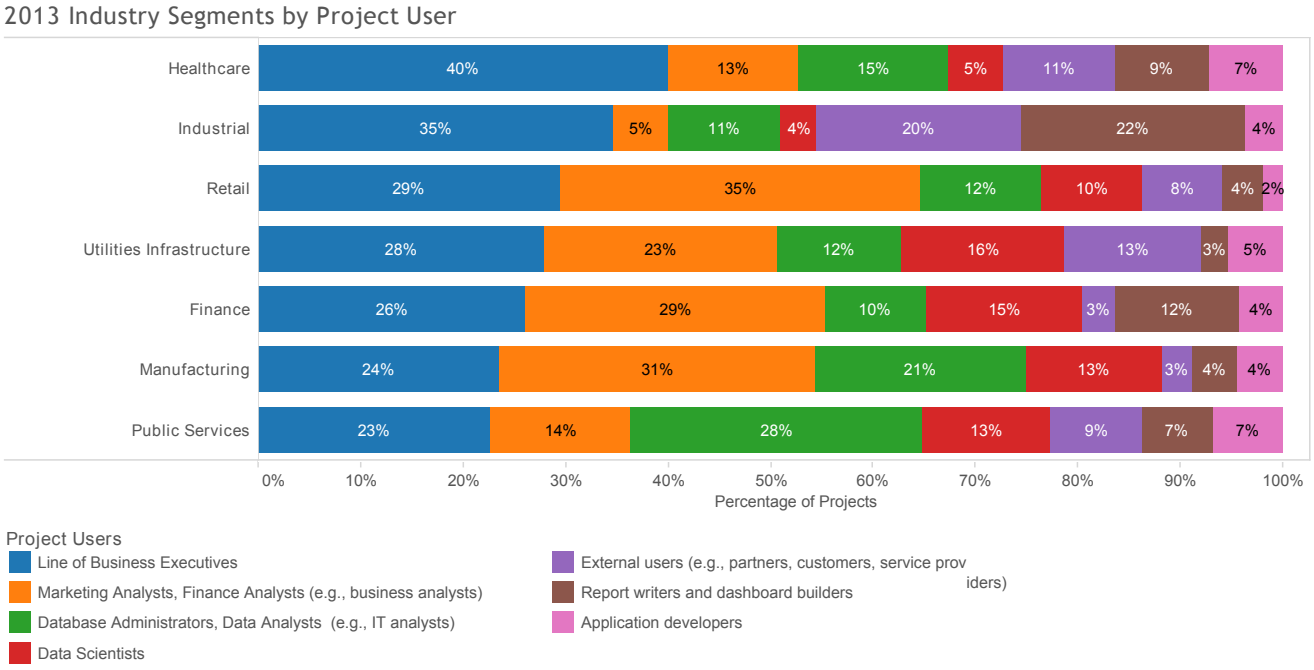


Figure 40

Examining the individual information user roles, there are some interesting observations. **Marketing and Finance Analysts** are the main users of Big Data projects in the **Retail** segment. This is supported by the increasing role of business stakeholders in using the output from Big Data projects. The role of **Data Scientist** is, however, negligible in the **Healthcare** and **Industrial** segments. **Data Scientists** appear to play a more active role in the other five industry segments in the 2013 research results.

When compared to the 2012 results, the highly specialized role of the Data Scientist has diminished in importance (see Figure 41 for comparison).

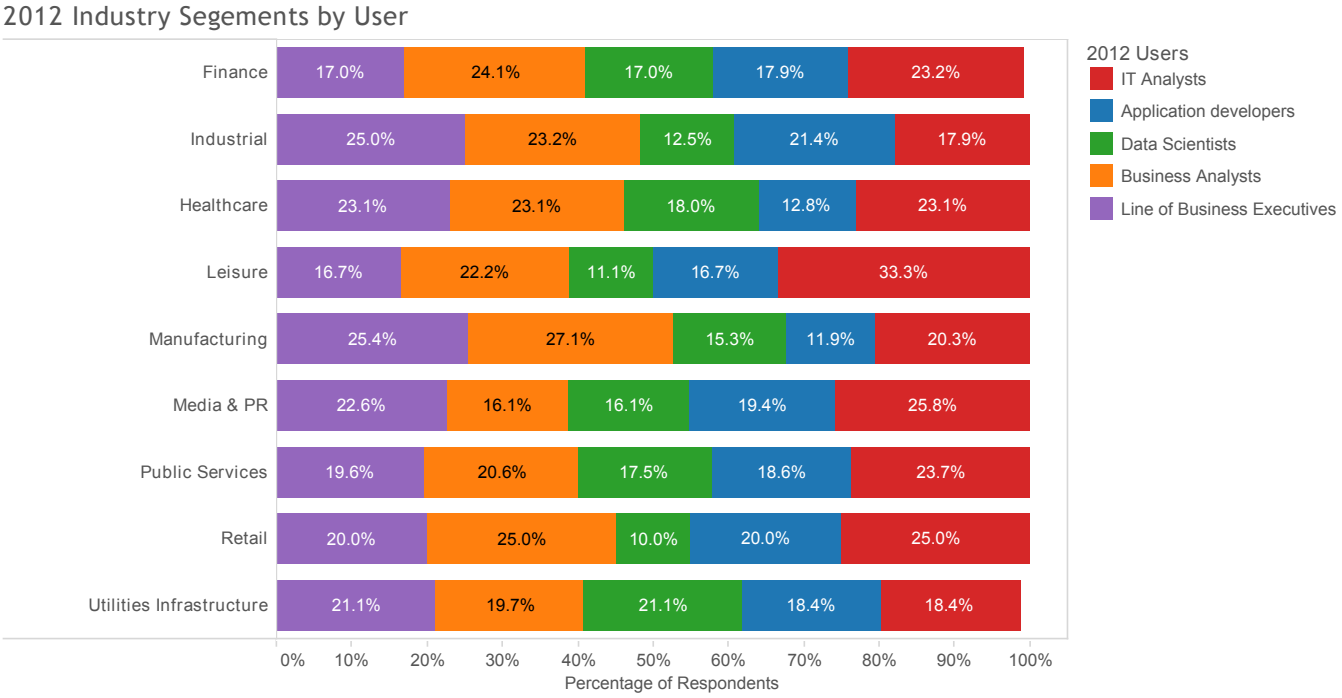


Figure 41

The role of **Application Developer** was substantial across all industry segments in 2012 ranking as high as 20% in the **Retail** segment. The project data from 2013 illustrates a shift to a much lower utilization by this user role. This may be explained by the inclusion of a new category in 2013 for **Report Writer** allowing 2013 respondents to be more specific in their responses.

5.3.2. Building Big Data User Skills

Leveraging the proper skills sets within an organization is critical to delivering a successful project. This is especially true with many Big Data initiatives. These projects often utilize multiple, new platforms. Of the 2013 respondents, 34% list skill gaps with multi-structured data platforms such as Hadoop or MongoDB as a significant challenge. In fact, overall skill development is a challenge across the data management landscape.

Leveraging the proper skills sets within an organization is critical to delivering a successful project. This is especially true with many Big Data initiatives.

Companies entering into a Big Data project are concerned about skill acquisition and development to drive Big Data projects. Many organizations, nearly 45%, are looking within their organization to **Organic Internal Staff Development** to fill this skills gap.

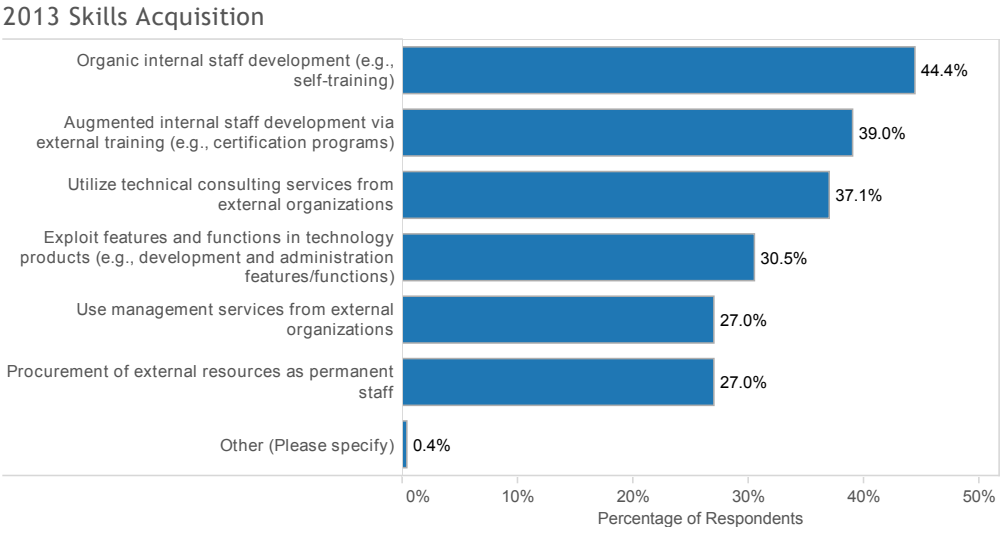


Figure 42

Almost four in ten (39%) are turning to **External Training via Certification Programs** for professional development. The Big Data hardware and software vendor community has identified this as an opportunity. Many Big Data vendors are focusing on education services and training to speed their client time to implementation. Not surprisingly, consulting organizations will benefit from the growth in Big Data projects. 37% of our respondents indicate that **Technical Consulting Services** will help bridge their skills gap.

Examining the challenge through industry segments shed light on how each will address the skill gap. **Healthcare**, **Retail** and **Public Services** industry segments leading strategy is to support **Organic Internal Staff Development**.

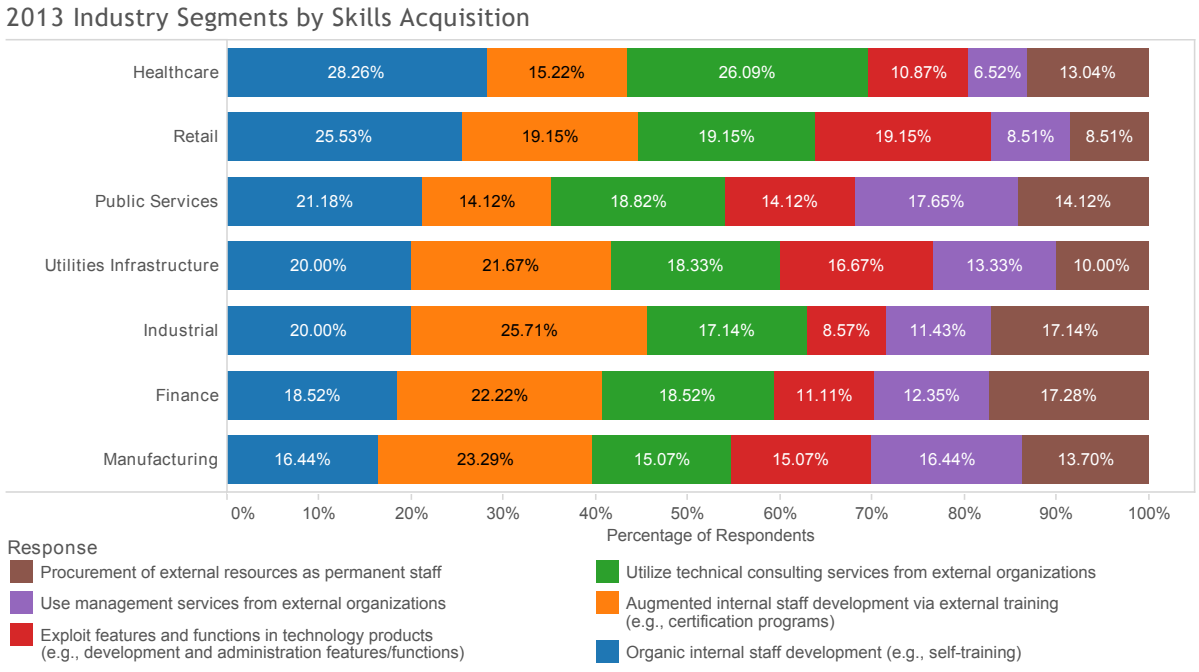


Figure 43

The **Industrial**, **Manufacturing** and **Finance** segments are more likely to focus on **External Training via Certification Programs** for internal staff to close the skills gap. **Industrial** and **Finance** are more likely to **Procure External Resources as Permanent Staff** to solve this issue than other industry segments.

5.4. Big Data Champions

All technology projects require sponsorship both in terms of IT support and budget funding. This is no different in the area of Big Data. Big Data projects are not trivial and the majority of projects require the acquisition of new hardware and software infrastructure.

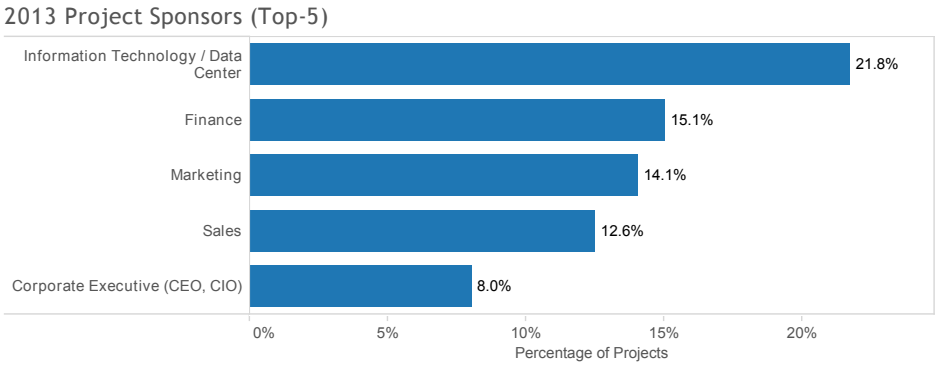


Figure 44

In many instances, Big Data project sponsorship comes from departments outside of the office of the CIO. This trend is another proof point in the maturity of the market. Diverse areas of the organization are embracing Big Data to solve critical business challenges. **Finance**, **Marketing** and **Sales** account for 41.8% of the sponsorship in the nearly 600 projects analyzed in this research

Interesting trends can be identified when comparing sponsors and industry segments. There is not parity across them. **Information Technology/Data Center** sponsor 38% of the Big Data projects in **Public Services** segment and 24% in the **Utilities Infrastructure** segment. The **Finance** department primarily sponsors projects in the **Retail**, **Healthcare** and **Manufacturing** segments.

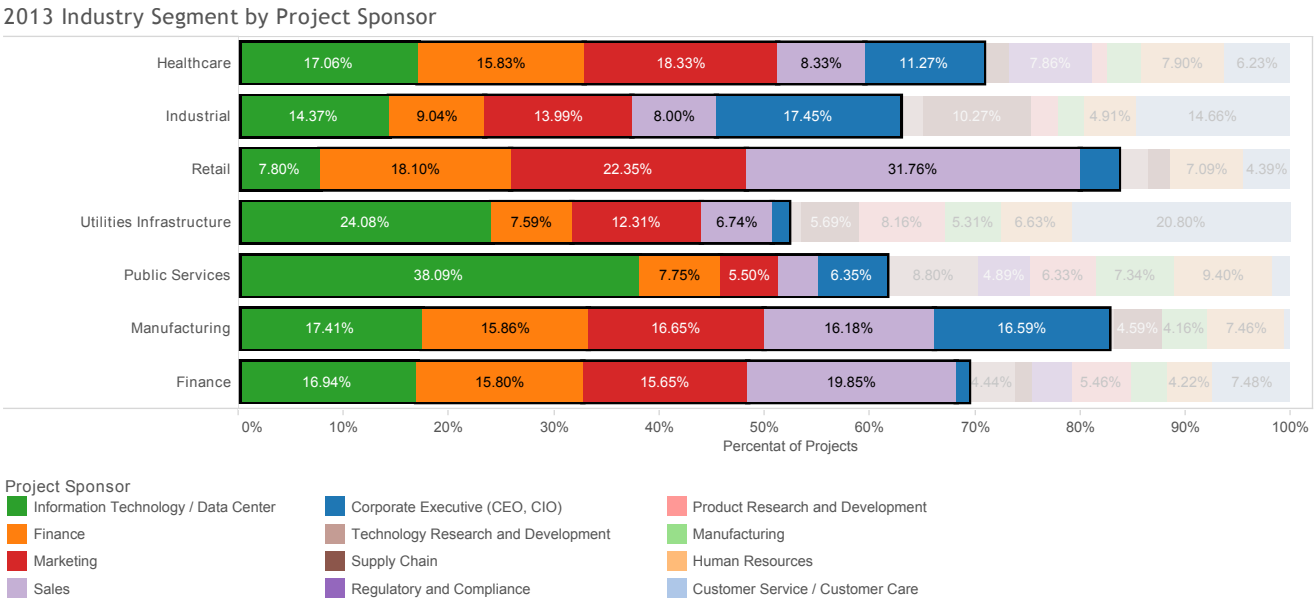


Figure 45

Marketing teams sponsor 22.4% of projects within the **Retail** space followed by **Healthcare** at 18.33% and Manufacturing at 16.7%. It is clear that a different mix of project sponsors drives each industry segment. This diversity is a result of growing adoption as well as growing success.

When comparing the 2012 respondents and 2013 project characteristic data, there are some noticeable trends. The project sponsor role of **Research and Development** (**R&D**) teams, including both Product R&D as well as Technology R&D, has shifted significantly.

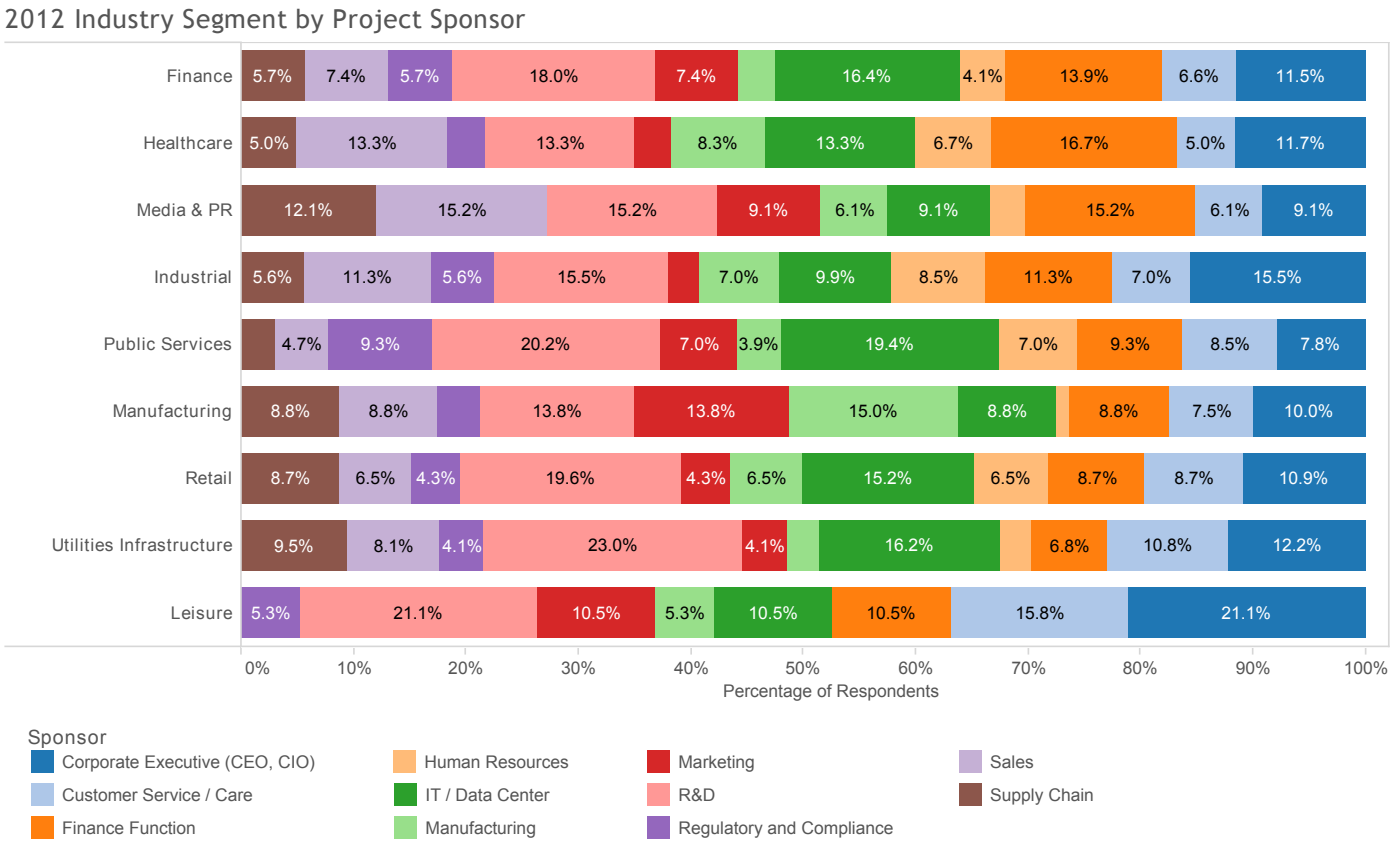


Figure 46

In 2012, **R&D** teams were a double-digit sponsor in all industry segments. When compared to 2013, the impact of this group on Big Data projects has shifted dramatically to other areas. It is clear that the office of CMO, and others in the CxO suite, are becoming a growing force in the commissioning and use of Big Data projects.

5.5. Building Blocks

As companies embrace Big Data, they need to determine how best to deploy Big Data projects. A decision must be made whether to leverage existing technology assets or to invest in new infrastructure. According to the 2013 respondents, 34.7% of Big Data projects were deployed using **Existing Hardware and Software**. This indicates Big Data projects will have a financial influence on budgets with 65.3% of projects acquiring new hardware and software infrastructure.

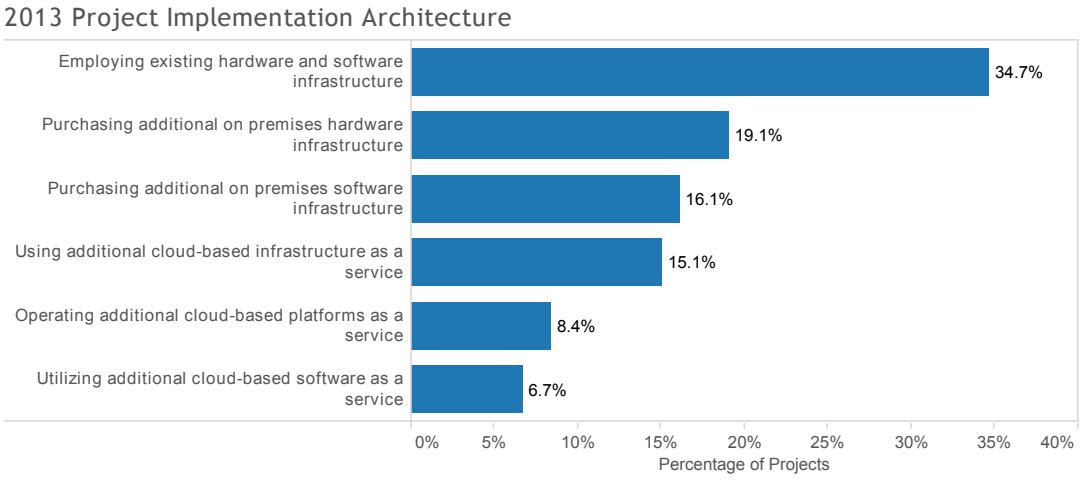


Figure 47

These purchases appear to be split between **Purchasing Additional On-Premises Hardware** (19.1%) and **Purchasing Additional On-Premises Software** (16.1%) Big Data projects are also making investments in cloud-based infrastructure. Of the 2013 Big Data projects, 30.2% of new Big Data project investments will be associated with Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS) offerings.

Implementation Architecture strategies vary across industry segments. Projects in the **Industrial**, **Retail**, **Finance** and **Healthcare** sectors deploy on **Existing Software and Hardware** greater than 40% of the time. Hardware investments are led by projects in the **Manufacturing** segment with 23.5% of projects implemented on new hardware.

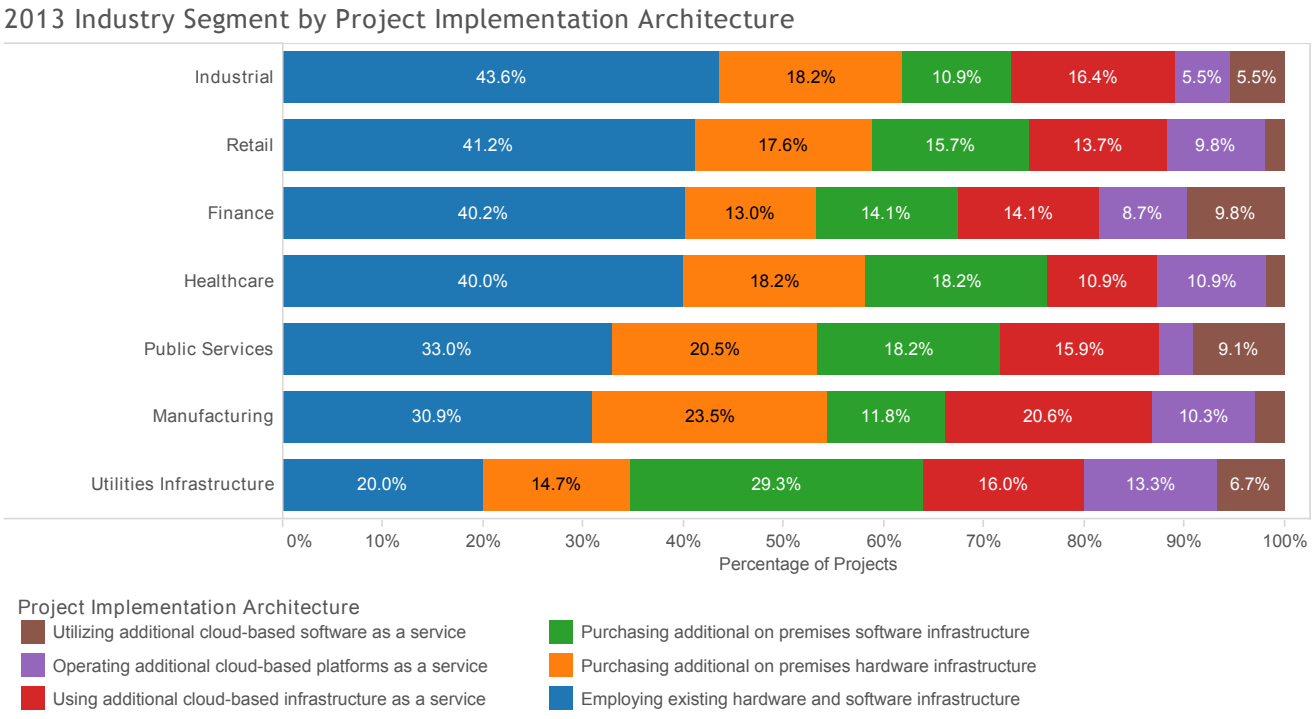
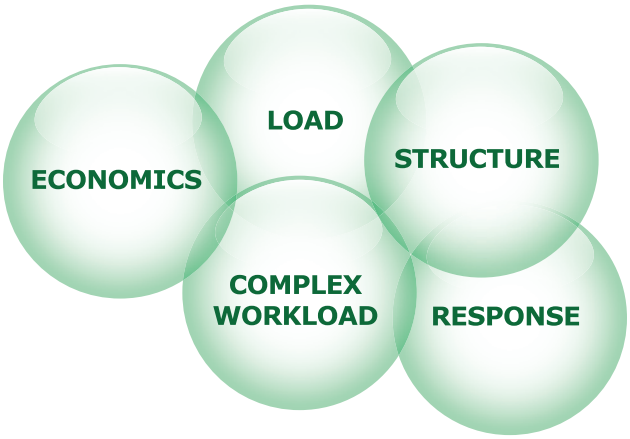


Figure 48

The **Utilities Infrastructure** sector invested in **New Software** for 29.3% of its Big Data projects. Cloud-based solutions are popular in all sectors. The **Utilities Infrastructure** industry segment is deploying 36% of their Big Data projects on newly cloud-based architecture. Cloud-based options are also popular with **Finance** sector. Of the activities in the **Finance** segment, 32.6% of projects utilize some form of cloud-base architecture.

6. Big Data Requirements

In 2012, the EMA/9sight research identified five core requirements of Big Data initiatives according to the indications of the panel respondents:



- ***Response***: New technology platforms such as Big Data tools and frameworks are at the core of this evolution and powering new solutions and improved speed of results.
- ***Economics***: Big Data platforms leverage commodity hardware, and the software is often free, substantially reducing the financial barriers to adoption.
- ***Workload***: Big Data platforms play a role within the ecosystem to execute extremely complex analytic workloads, and innovative companies are willing to invest early in these solutions to gain competitive advantage.
- ***Load***: Data loads are growing more complex and the sources are more diverse. Driven by greater complexity and demand, Big Data adoption is driven by the need to provide flexibility.
- ***Structure***: Data structure and schema flexibility is key to the foundation of Big Data utilization and adoption.

6.1. The Need for Speed

The requirement to provide faster processing response is core to the revolution in Big Data. It is not sufficient to provide processing in Big Data environments in a fashion consistent with traditional batch processing, supporting the premise that Big Data requires platforms that go beyond Hadoop. The results, whether they are operational or analytical, need to show improvement over the speeds associated with existing or traditional platforms. The expectation is that regardless of the processing being requested the results will be faster than previous eras of computing.

6.1.1. Building the Use Case for Speed

When EMA/9sight examined the use cases of the respondents above, the **Speed of Processing Response** was the top choice among respondents.

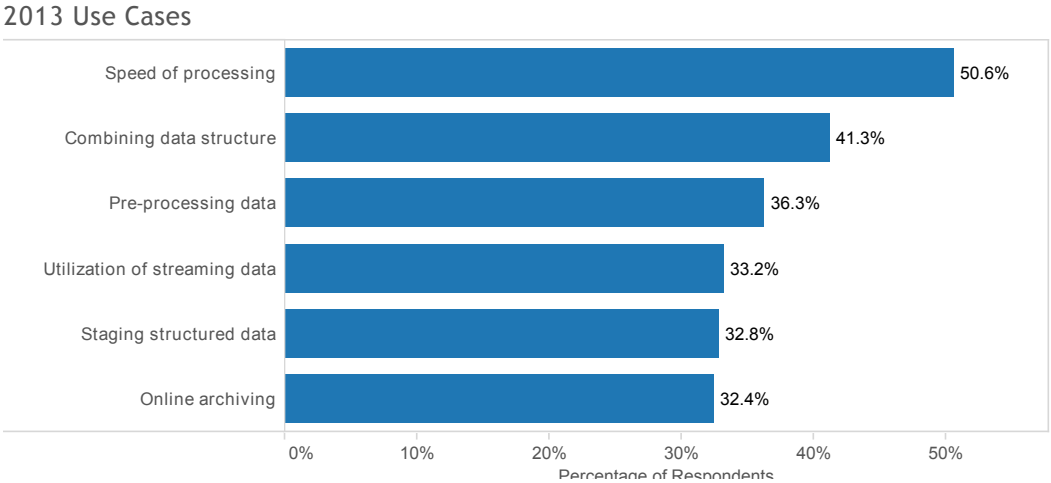


Figure 49

This shows how important **Response** is when considering the requirements associated with Big Data initiatives. It is reflected not only in the use cases of the EMA/9sight survey respondents, but also in how they are implementing their projects as this report detailed earlier.

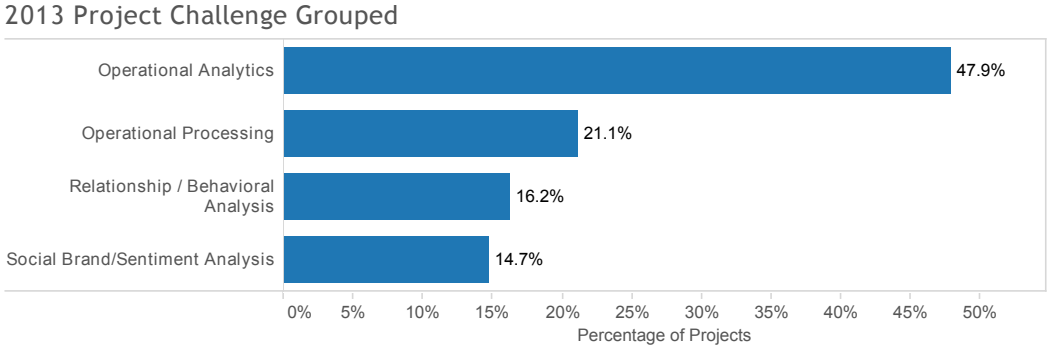


Figure 50

The key to many of the **Operational Analytical** projects is to provide a speed of **Response** consistent with the speed required by their operational workflows. For example, speed of processing is important in industries such as Telecommunications and Online Consumer Retail. In these industries, the businesses of intangible goods and services have low barriers to entry and exit. Fraud management and cross-sell/up-sell projects can provide key elements to competitive advantage as well as profit and loss. This comes not just as a technical initiative, but also as a clearly identifiable business requirement.

The Big Data projects of the 2013 EMA/9sight respondents show that **Speed of Processing Response** use cases have the second most projects **In Operation**. This indicates that the importance of **Response** is not just a strategy, but is shown in implementation as well.

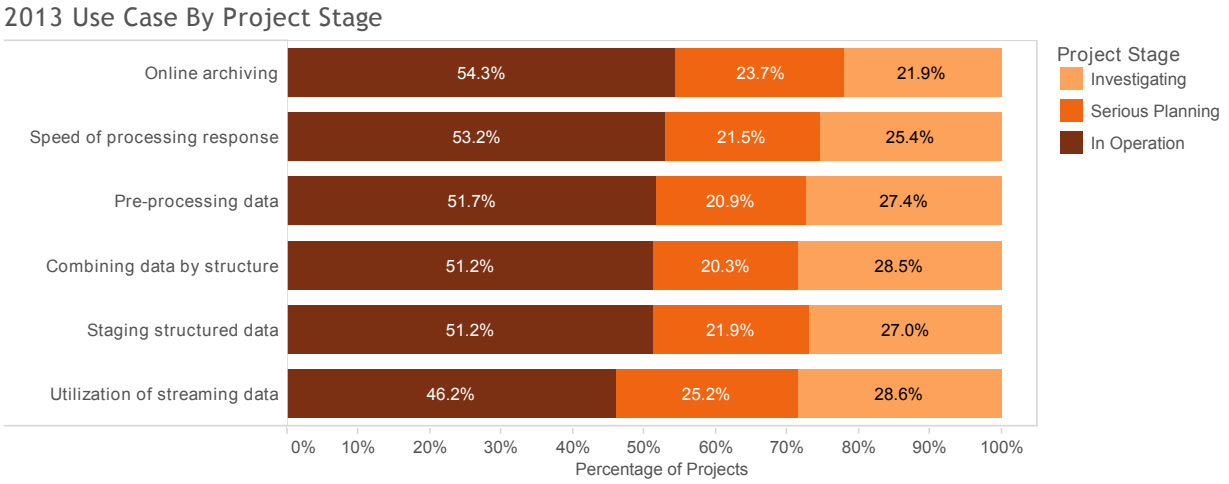


Figure 51

It is worth noting that the top use case from the 2012 survey was **Online Archiving**. Respondents with projects **In Operation** in 2013 are showing that last year's top use case strategy is also well represented in **In Operation** projects.

6.1.2. Technical Drivers Motivating Response

There are times when IT departments will strive for technical solutions when the business requirements do not support the effort. With Response, this is not the case. While **Scaling Issues with Current Platform** is the top response from the EMA/9sight panel respondents, the speed of **Response** technical drivers are the second and third highest responses.

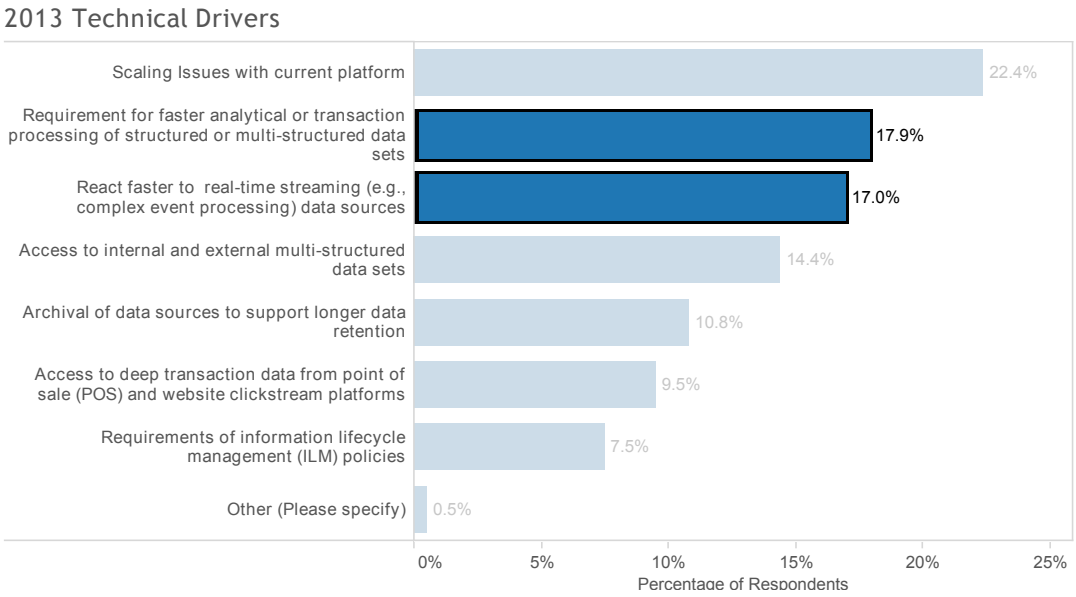


Figure 52

Whether it is for faster processing or faster access to streaming data sources, the technical drivers of the 2013 EMA/9sight panel respondents support the Big Data requirement for high rates of Response associated with their Big Data initiatives.

6.2. No Such Thing as a Free Lunch

The economics of particular Big Data technologies and open source alternatives has been described as: ***"Free… like a free puppy"***

This indicates that while you might be able to obtain the software inexpensively; the hardware and headcount support costs need to be included in any budgeting exercise. Yet, alternatives to traditional data management solutions are pressing down on the overall costs associated with Big Data initiatives. As you can see in Figure 53, among EMA/9sight respondents, **Improved Data Management and Competitive Advantage** is the top Business Challenge for their Big Data Initiative.

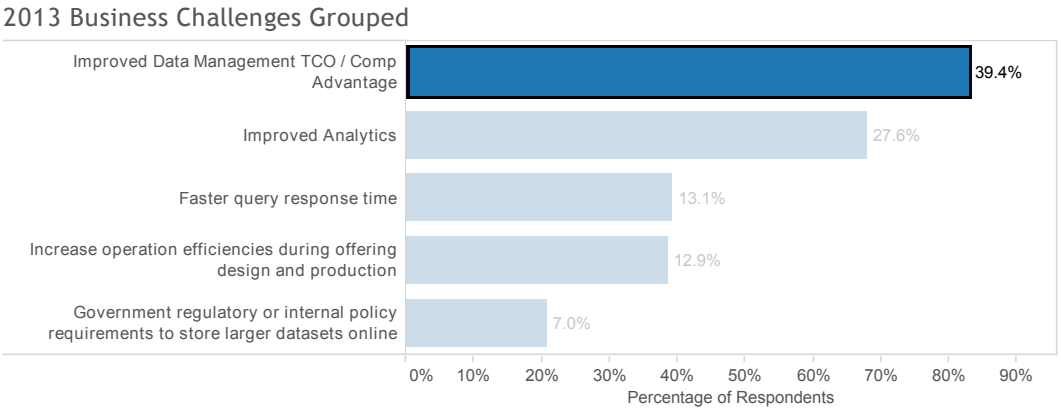


Figure 53

It is this business imperative and the downward pressure on pricing derived from the perception that certain Big Data technologies are "free" that drives expectations that organizations should do more with less. In that vein, EMA/9sight panel respondents were asked to identify their overall IT and Big Data budgets.

6.2.1. Looking at Information Technology Budget

The 2013 EMA/9sight respondents were surveyed in regards to their overall annual information technology budget, the purpose of which was to identify an approximate Big Data budget for 2013.

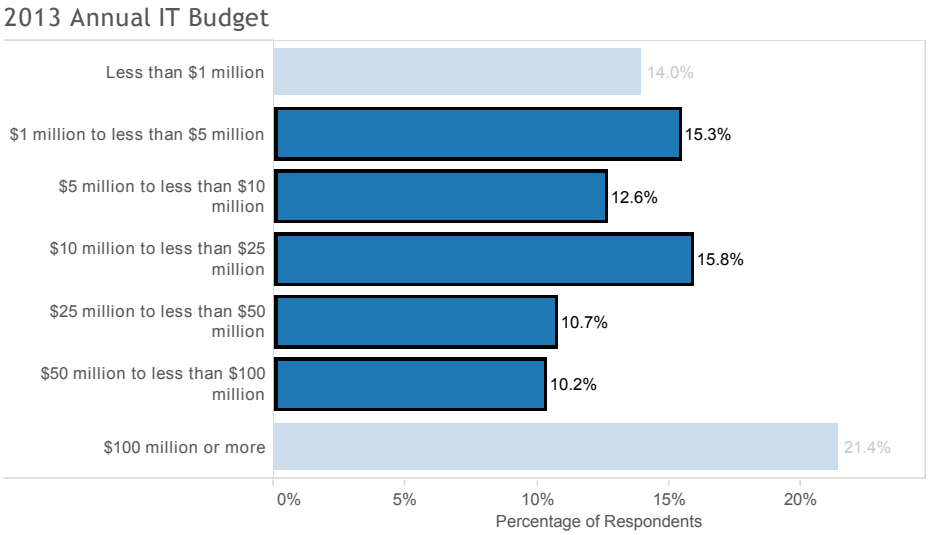


Figure 54

Based on the information in Figure 54, nearly 65% of respondents indicated that their annual IT budgets were between $1m USD and $100m USD. This "supermajority" answer represents the significant portion of the EMA/9sight respondents' indications as to their budget. When you look at the "core" or most common responses from the EMA/9sight panel below, you find that nearly 30% of respondents indicated a budget between $5m-$25m for 2013.

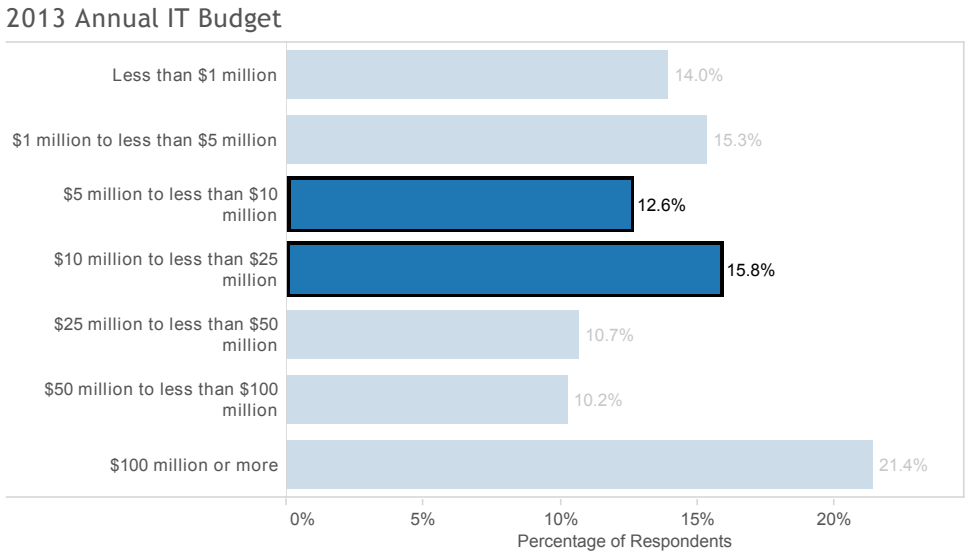


Figure 55

It should be noted that the single largest respondent category is in the $100m and above category. While these respondents are not being discounted, their inclusion as the most common category would shift the budget calculations by 4-20x.

When asked to indicate what percentage of their overall IT budget was assigned to their organization's Big Data initiative, the EMA/9sight panel respondents provided the following answers.

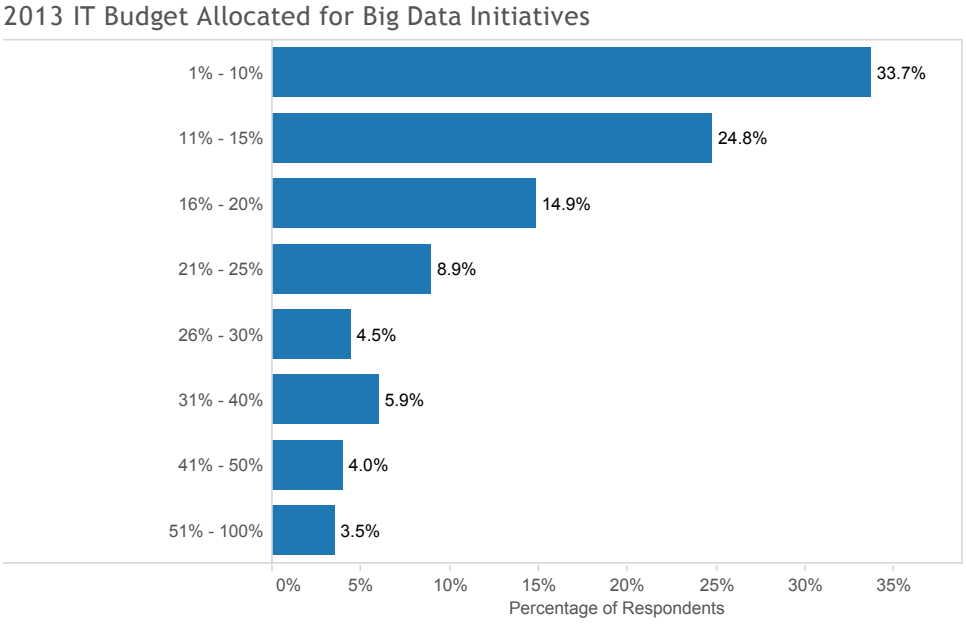


Figure 56

Over eight of ten responses show that between 1–25% of their 2013 IT Budget was assigned to their Big Data initiatives. When reviewing the single largest set of responses, it is clear that nearly four of ten responses are associated with 11–20% of IT budget being assigned to Big Data initiatives.

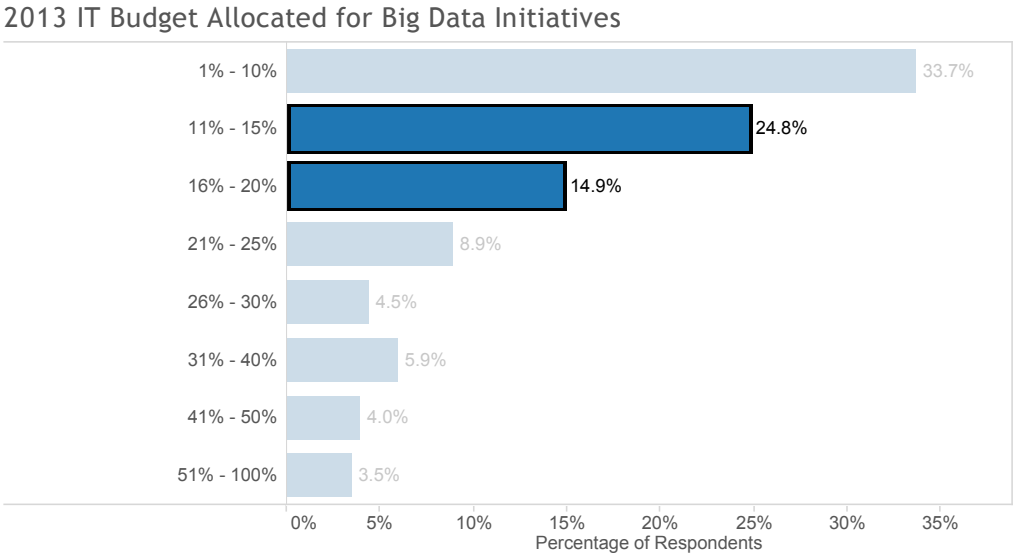


Figure 57

Again, it should be noted that the single largest response category was on one of the ends of the response spectrum. This single category is not being ignored, but rather considered a set of responses that may overly influence the Big Data budgeting calculations by 2-20x.

6.2.2. Projecting Budgets and Allocations

From this information linking IT Budgets with estimated Big Data expenditures, the calculation can be made that for "supermajority" budgets associated with Big Data initiatives across all industries and company sizes for 2013, the budget range would be between $10,000 and $20m USD.[14] A more typical or "common" Big Data budget for 2013 would be represented by $550,000 to $5m USD.[15]

Since the sponsors of Big Data initiatives are spread across a majority of EMA/9sight responding organizations, it makes sense that organizations beyond the CIO's office would be responsible in some part for Big Data projects and programs.

When asked if there were contributions to Big Data budgets from outside of the overall IT budget, fewer than 50% of respondents indicated that this was the case.

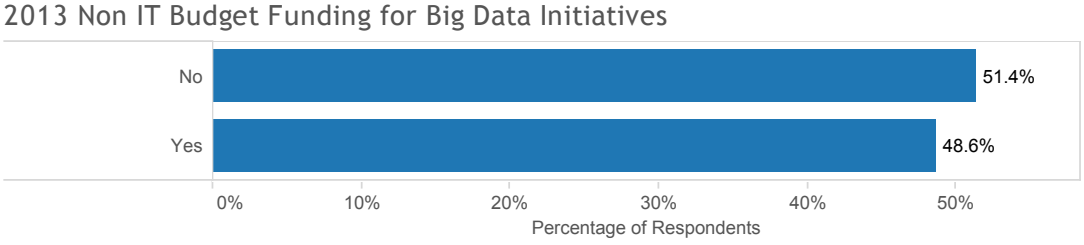


Figure 58

[14] $1m USD – $100m USD annual IT Budget x 1%-20% Big Data allocation = $10,000 - $20m budget range
[15] $5m USD – $25m USD annual IT Budget x 11%-20% Big Data allocation = $550,000 - $5m budget range

Based on respondents who indicated funding stakeholders beyond the CIO and who have knowledge of their organization's budget distributions, a "supermajority" of respondents indicated that between 11% and 50% of Big Data initiatives were being funded by those non-CIO stakeholders.

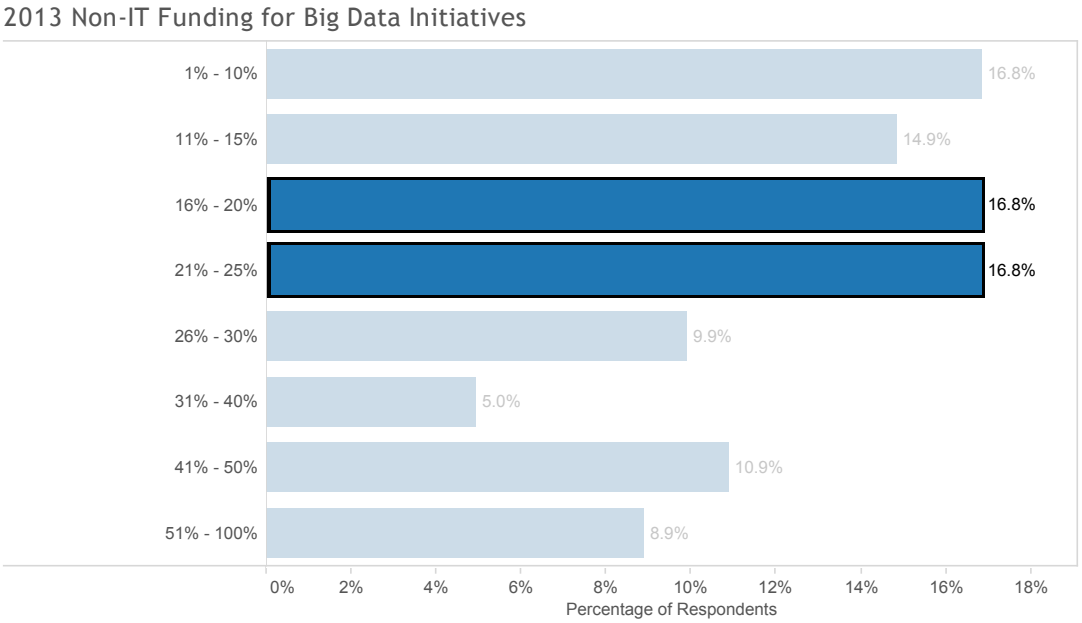


Figure 59

The most common set of answers indicated that non-IT based resources were providing 16–25% of those Big Data budgets. This results in between $1,100 at the very lowest end to $10m of Big Data funding resources coming from sources outside the CIO's office.[16] The most common responses indicate that nearly $100,000 annually to just over $1.2m[17] are coming from budgets associated with **Finance**, **Marketing** and **Sales** organizations.

6.2.3. The Average Budget

Using the information from the most common Big Data budget calculated above and the most common indication number of projects associated with EMA/9sight respondents, you see that budgets are being allocated between $275,000 and $2.5m for a single project.[18] When you include software license and maintenance, hardware purchase and maintenance, and headcount to support these environments, you can see that overall **Economic** cost pressures will continue to impact how Big Data initiatives are implemented in 2013 and moving into 2014.

6.3. Complex Workloads Go Real Time

Complex processing for operational and analytical workloads is important for Big Data initiatives. No longer are "simple analytics" associated with sum totals or simple queries acceptable. Business stakeholders want the ability to include complex business rules and advanced analytics in the form of predictive models and natural language processing added into both their analytical request and their operational workflows. This increase in the complexity of processing workload is driving Big Data initiatives to fulfill business requirements to develop and maintain competitive advantage.

[16] $10,000 USD – $20m USD annual IT Budget x 11%-50% Non-IT Allocation = $1,100 - $10m budget range
[17] $550,000 USD – $5m USD annual IT Budget x 16%-25% Non-IT Allocation = $88,000 - $1,250,000 budget range
[18] $550,000 - $5m budget range / 2 projects = $275,000 - $2.5m per project budget

6.3.1. How Strategy Impacts Complexity

Processing workloads are different based on a chosen implementation strategy. In the EMA/9sight survey respondents were asked to identify the implementation strategy associated with their Big Data initiatives. They were given the opportunity to describe their implementation strategy as **Operational**, **Analytical** or **Exploratory** in nature.

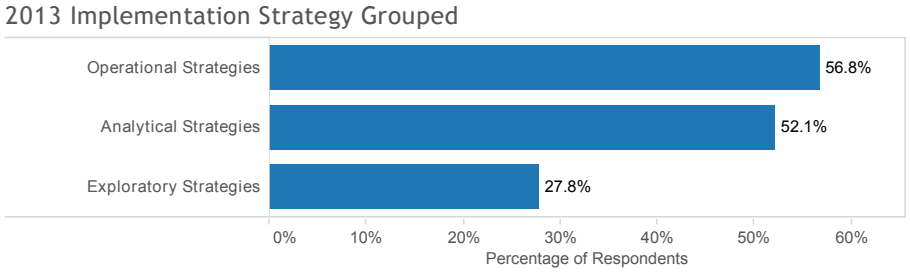


Figure 60

Nearly 60% of respondents indicated that they were using an **Operational** strategy for Big Data. These approaches include using **Machine-generated** datasets to improve efficiency and increase sales. Over 50% indicated that they were using Big Data initiatives associated with some type of **Analytical** strategy. This approach includes the ability to link multiple datasets such as **Human-sourced** information and **Process-mediated** data sources to more closely identify customer or potential customer behavior. Almost 30% indicated that they were using an ad-hoc **Exploration** approach. This approach would include using a combination of all three Big Data domains.

Looking deeper into the industry segments using various implementation strategies, **Healthcare** and **Public Services** are the top two segments associated with the **Operational** strategy. This comes from the heavy process-driven approaches that both of those segments employ. Risk mitigation and process compliance are both heavily involved in the operational models of those two industry segments.

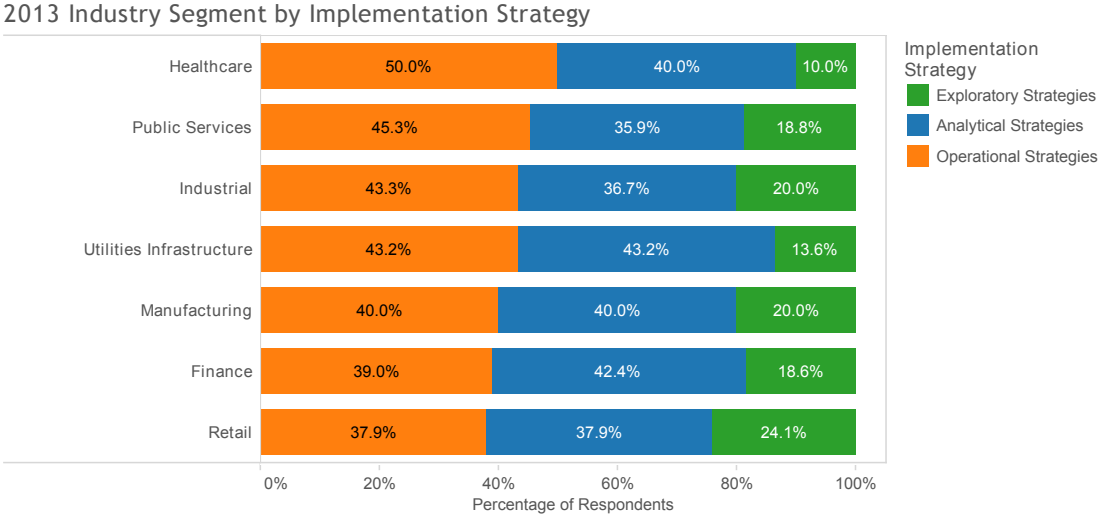


Figure 61

Utilities Infrastructure and **Finance** show the largest indications of an **Analytical** strategy of the industry segments. Finance and Banking have the most history associated with **Analytical** strategies. Utilities again are finding new avenues for analytics. This is driven by both constraints on power generation and natural resource availability.

6.3.2. Business Drivers of Complex Workload

When asked what their top three business drivers associated with Big Data were, the EMA/9sight respondents indicated that **Improved Analytics** was second only to the **Improved Data Management TCO / Competitive Advantage** response mentioned in the **Economics** segment. This confirms the implementation strategy analysis above for the **Analytical** approach.

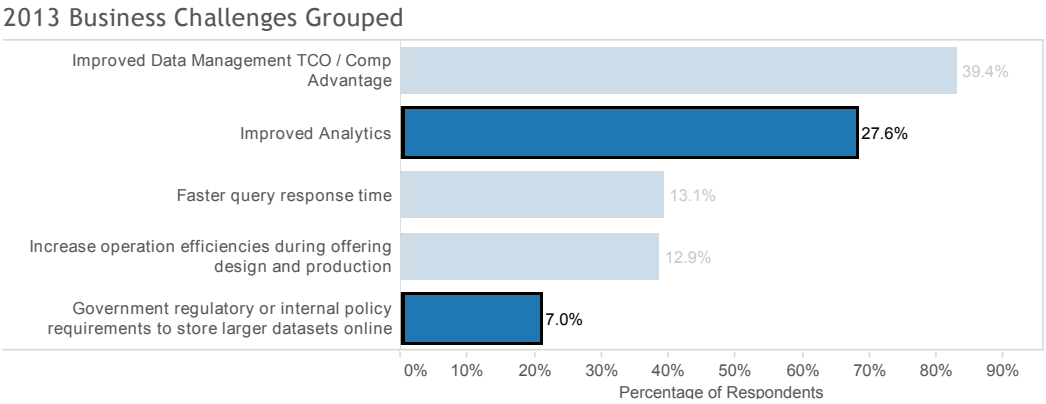


Figure 62

While it was not the top answer, the business driver of using Big Data initiatives to solve **Government Regulatory or Internal Policy Requirements** supports the analysis associated with the Operational strategies.

In fact, when detailing the breakout of industry segments, **Healthcare** and **Public Services** again are strongly associated with **Government Regulatory or Internal Policy Requirements** as a business driver for their Big Data initiatives.

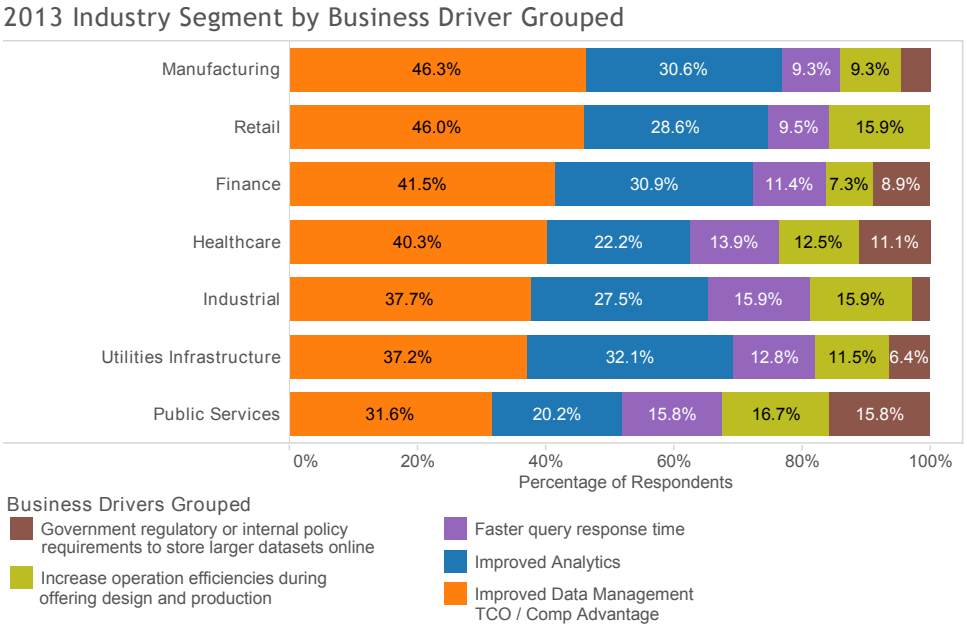


Figure 63

In the 2013 survey, **Utilities Infrastructure**, **Finance** and **Manufacturing** industry segment all focused on Improved Analytics as part of their top-3 business drivers.

6.3.3. Complex Challenges across Industry

Returning to the business challenges associated with individual Big Data projects and additional confirmation of the use of complex **Workload**s across industry segments; **Industrial**, **Public Services** and **Utilities Infrastructure** are focused on **Operational Analytics**-based projects. These projects put considerable pressure on the workload capabilities of the underlying platforms because of the complexity of the analytics and the requirement to integrate the results in realtime into operational processes.

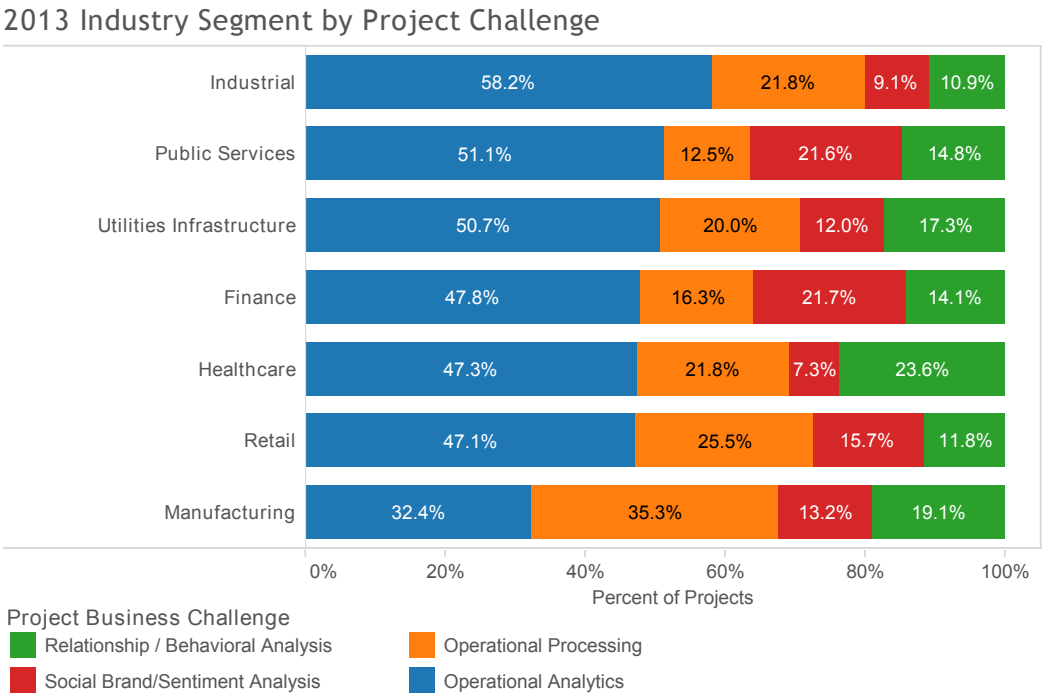
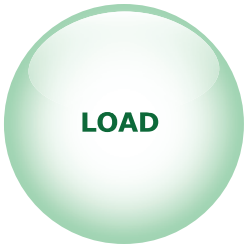


Figure 64

As you move to the more advanced analytical projects, the Finance industry segment is a natural fit for **Relationship / Behavioral Analysis** and **Social Brand and Sentiment Analysis**. However, a less obvious candidate, **Public Services**, is the top industry group for those two response groups. It will be interesting to see how public sector organizations utilize the results from those projects to meet the needs and requirements of their local and national constituents.

6.4. Data Loads Get Bigger…and Smaller



Big Data is called "big" for a reason. Many of the datasets associated with Big Data initiatives are enlarging the overall data management environment for companies. In 2012, the EMA/9sight Big Data survey sought to quantify the size of the data stored within various operational or analytical Big Data Platforms. In other words, to define the "big" of Big Data in terms of data under management. EMA/9sight has again, in 2013, worked to provide an estimate of the Big Data portion of data management environments.

6.4.1. Overall Environment Sizing

When respondents were surveyed in regards to their overall data management environment, nearly 75% of the EMA/9sight survey respondents indicated that they were managing between 10TB and 5PB of data. This represents a significant shift from 2012 survey respondents.

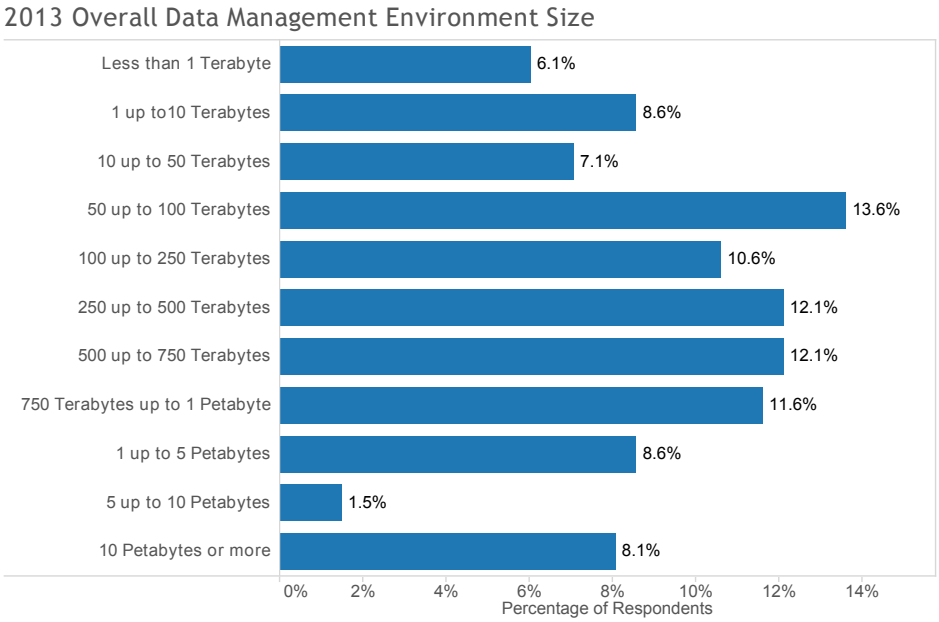


Figure 65

A significantly larger percentage of the panel respondents in the 2013 EMA/9sight survey indicated that they were managing over 10PB of data as part of their data management environment. This increased the range of environment sizing significantly. The 2013 low end is 10x larger (10TB in 2013 vs. 1TB in 2012) than 2012. The 2013 high end is nearly 7x larger (5PB in 2013 vs. 750TB in 2012).

The most common range for 2013 of just over 36% respondents is between 50TB–500TB. This range is also significantly different from the 2012 survey. The 2012 most common range was 50TB–100TB.

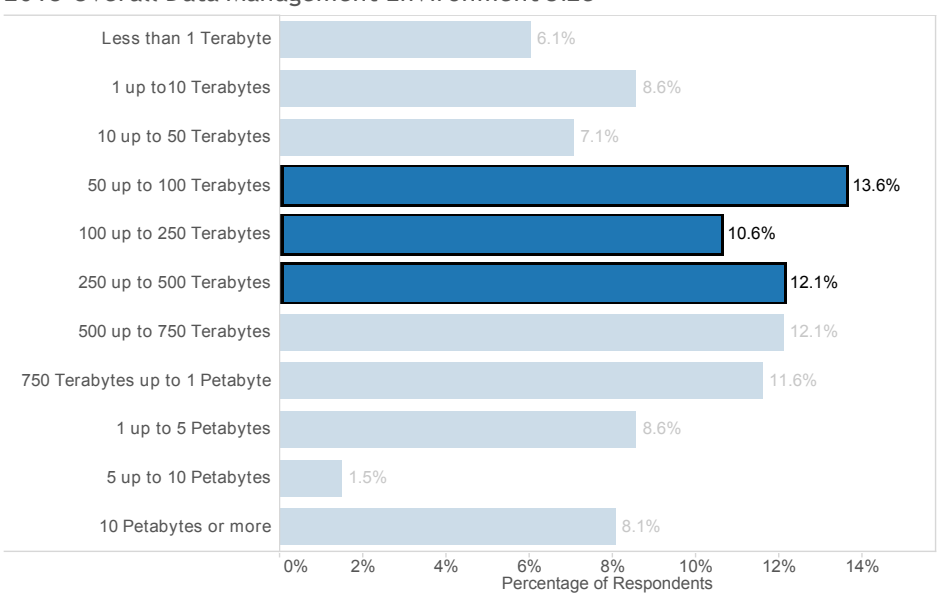


Figure 66

The contingent of significantly larger environments in the 2013 study may be influencing this increase.

6.4.2. Sizing Big Data Environments in 2013

When asked what percentage of their overall environments was dedicated to Big Data related data management, over seven of ten respondents indicated that they were using between 11% and 40% of their data management environment on Big Data solution. This "supermajority" range of respondents was similar to the responses in 2012.

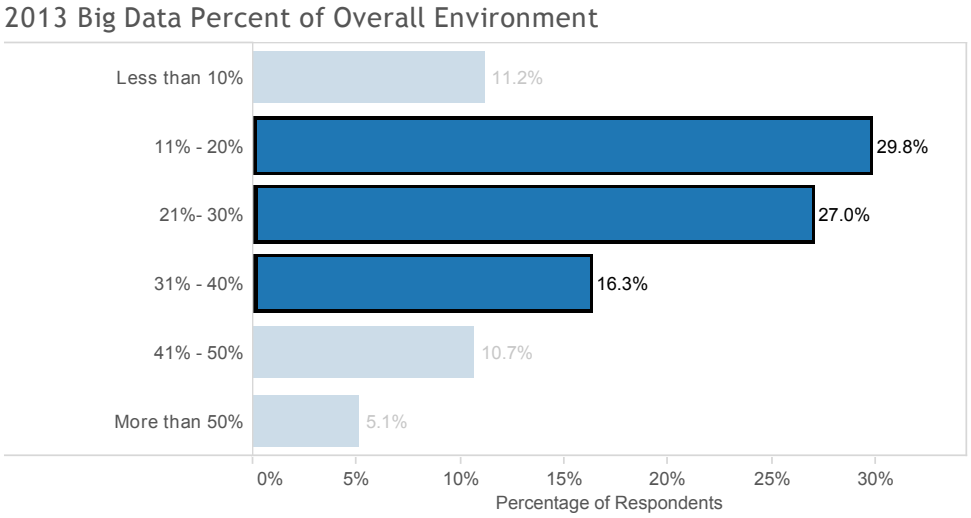


Figure 67

The most common indication among the panel respondents was in the 11–20% range of data management environments associated with Big Data initiatives. With the most common response in 2012 being 21–30%, the most common response in 2013 response represents a lower percentage of the overall data management environment.

Using the information from the EMA/9sight respondents on data management environment size and percentage dedicated to Big Data initiatives, the most "common" Big Data solution environment can be sized to approximately 5.5TB and 100TB.[19] When the larger range of data management environment sizing is considered Big Data environments range can span 10TB and 2PB[20].

The most common range represents a wider spread of Big Data environment sizes than the 2012 calculated size. On the low end, 2013 represents an approximately 50% reduction from last year (5.5TB vs. 10TB). The high end for 2013 is an over 3x increase (100TB vs. 30TB). For the "supermajority" range, the calculated ranges represent an associated increase in the size of Big Data environments with the increase of the base data management environment. The 2013 "supermajority" range of 1.1TB to 2PB represents a 10x increase for the lower range and a nearly 7x increase for the upper range.

When looking at how the 2013 "calculated projection" compares with this year's survey "actual," the 2012 numbers represent an approximately 2.5x increase for most projections from just one year ago. This seems to indicate that the EMA/9sight panel respondents were experiencing even larger growth than they predicted with their estimates last year.

[19] 50TB-500TB x 11%-20% = 5.5TB – 100TB size range
[20] 10TB-5PB x 11%-40% = 1.1TB – 2PB size range

6.4.3. Projecting Data Loads in 2014

Looking to next year in 2014, EMA/9sight respondents indicated how their Big Data environments would grow in the next year.

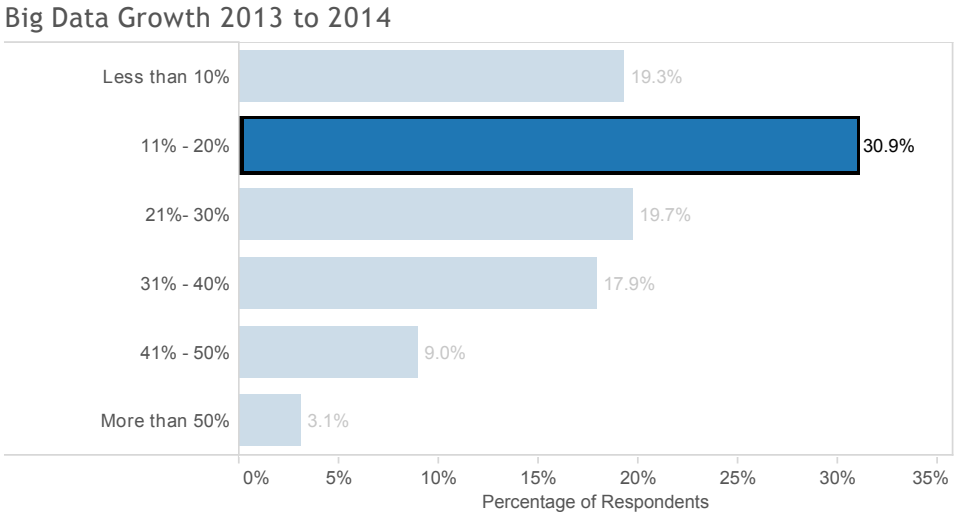


Figure 68

When asked to project growth for 2014, 69% of respondents said that indicated growth rate for Big Data was between 11% and 40%. This is a similar rate of change to the 2012 EMA/9sight survey respondents. The most common response in 2013 was in the 11–20% range. Again, this is a reduction from the 2012 most common response of 21–30%.

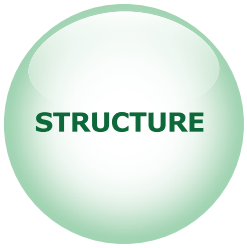
Using these answers in association with the 2013 Big Data environment sizing above, an approximation of the most common (27%) Big Data solution environment can be sized to approximately 6TB to 120TB.[21] When the wider range of responses is used, the environment range moves to 1.2TB to 2.8PB.[22]

The "supermajority" range also represents a similar growth from 2012, but with a related increase for the increased size of the overall data management environment. The "most common" 2014 projected environment also represents the wider range of environments. Both show that the implementers of Big Data solutions are interested in a wide spread of data sizes and not focused on simply the extremely large implementations.

When looking at the more common answers, there is a continued trend of multi-terabyte environments to now breaking the 100TB range. This shows that the most common environments are well above what can be easily handled by shadow IT departments and "skunk works" projects. These data **Load**s require a more professional and specialized data management environment, whether that is Hadoop, NoSQL or a standard SQL-based database management (DBMS) platform.

[21] 5.5TB – 100TB x 11%-20% increase = 6.1TB - 120TB size range
[22] 1.1TB – 2PB x 11% - 40% increase = 1.2GB – 2.8PB size range

6.5. Big Data: Un-Structured vs Multi-Structured



The term Big Data has done an excellent marketing job of delineating the difference between traditional relational data structures and this new era of multi-structured datasets. While some have described Big Data as being "un-structured," Big Data environments are comprised of data that comes in many structures and formats. From existing relational structured datasets to images and audio files, and vastly different types of information, these data sources each have their own specific data structures, if not data formats.

6.5.1. Breaking Down Big Data Domains

The information that is used in various Big Data environments comes from many sources in an enterprise. As stated above, the high-level categories of **Process-mediated**, **Machine-generated** and **Human-sourced** information provide high-level guidance. Of the 12 different data types, represented below, you can see that five are of the **Machine-generated** category including **Application and Server Log information** and **Network Activity and Probe Information**. Both of these data types represent significant contributions to Big Data environments, as they are the backbone of concepts such as the smart utilities grids and effective management of advanced networks to provide "burstable" access to cloud-based environments as well as between various internal and external data centers. The next two types, tied for third most indicated: Geo-location information from mobile devices and telematics, and Click-stream information from online application and mobile apps also follow this variable structure.

The term Big Data has done an excellent marketing job of delineating the difference between traditional relational data structures and this new era of multi-structured datasets. While some have described Big Data as being "un-structured," Big Data environments are comprised of data that comes in many structures and formats.

Each of these **Machine-generated** datasets is based on formatting standards as opposed to a standard format such as a relational structure. Examples of these formatting standards can be seen in traditional XML formats and as newer JSON formats. These formatting standards are best described as multi-structured since they have a single format, but may have multiple valid structures.

The top three **Process-mediated** data types associated with Big Data environments are **Operational Application Data** (e.g., point of sale, customer care, supply chain), **External Augmentation Data** (e.g., demographic or psychographic), and **Curated Business Information** (e.g., single version of truth customer or product information). Each of these data types generally follows a relational structured format.

Operational Application Data sources often follow a normalized format made popular by the work of Codd-Boyce, and appropriate for operational workloads minimizing database updates. External demographic and psychographic information as well as **Curated Business Information** generally follow a denormalized format appropriate for reporting and lookup workloads.

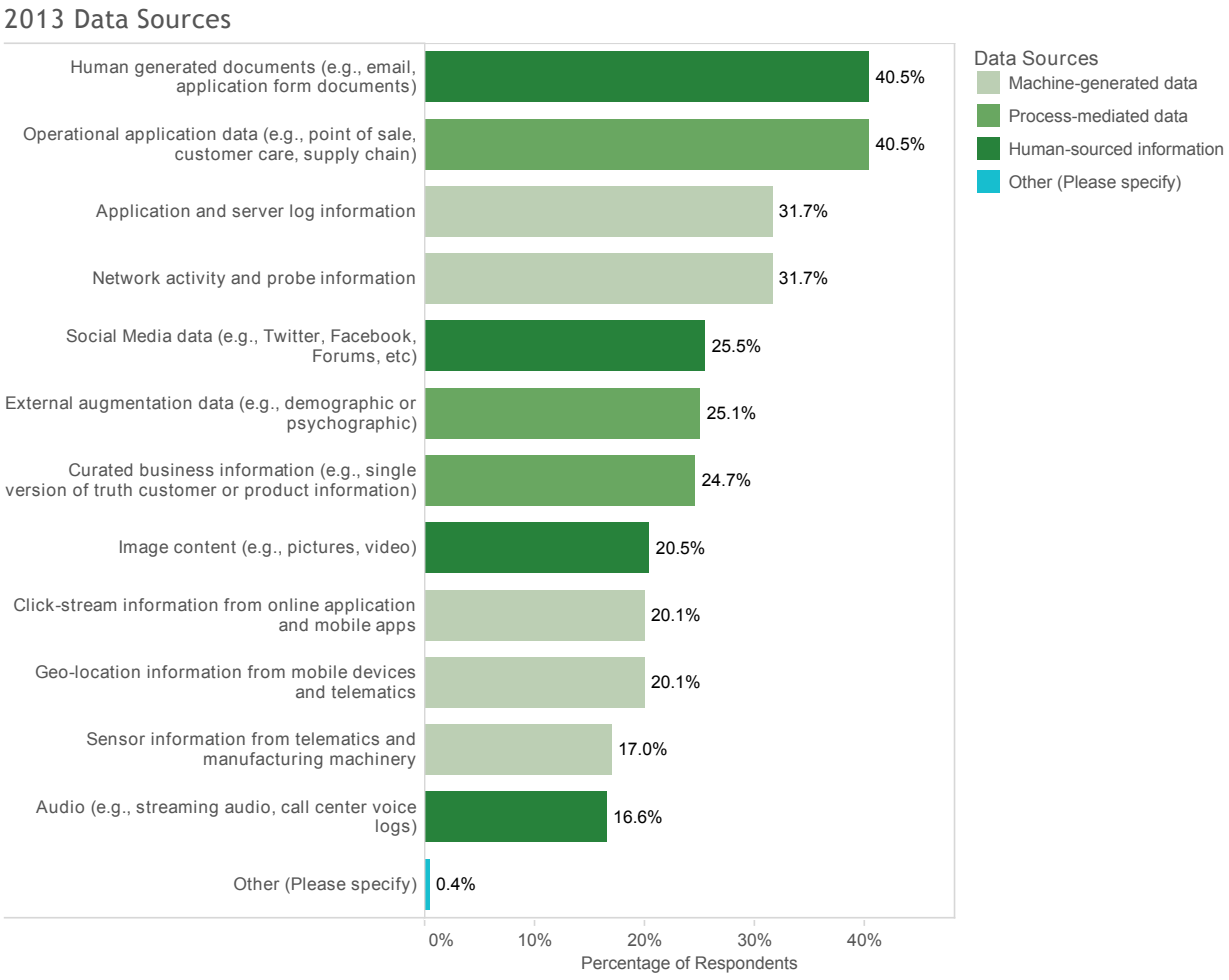


Figure 69

Human-sourced information in the form of documents is the highest indicated answer for EMA/9sight respondents. When you link this dataset with the next two highest selections of Social Media data (e.g. Twitter, Facebook, forums, etc.) and Image content (e.g., pictures, video), you begin to see the reason why so many describe Big Data as "unstructured data." Scanned forms and hand-written documents often look like one large blob of information appropriate more for a person to read or recognize than for a computer or automated system to analyze.

With the introduction of audio and video information into both private and public sector institutions, there is a rise of creativity to find the structure in those unstructured datasets. Natural Language Processing (NLP) and textual sentiment analysis can take handwritten notes and turn them into data appropriate for automated analysis. Images once considered the domain of individuals to categorize can now be analyzed to identify not only shapes and faces in a gross manner, but now those same images can be analyzed to identify specific objects based on the "structure" present in the image. Objects can be identified by size and color. Individuals can be identified by polygon structures associated with the distance between eyes, nose and chin.

6.5.2. Moving Big Data

To get information from the wide variety of data sources into a Big Data environment requires a considerable amount of coordination. When operational and analytical systems were integrating data between relational-based data structures, the integration process was relatively simple. With the introduction of multi-structured datasets and image/audio formats, the integration process becomes more complex. Multi-structured datasets can be integrated into relational formats, but you often end up with a sparsely populated table structure when multiple tags or fields are not consistent across all "records" in a format such as XML. Image/audio formats can be similarly integrated into relational formats, but again the relational structure is often forced or ill-suited for the originating dataset.

When asked to designate which data integration techniques are used in managing data within their Big Data environments, the 2013 EMA/9sight survey respondents indicated that their top three strategies were **Batch, bulk data integration**; **Real-time streaming data integration** and **Data replication (e.g. standardized duplication of enterprise data sources)**. Each of these techniques can be used with any of the above indicated data structures.

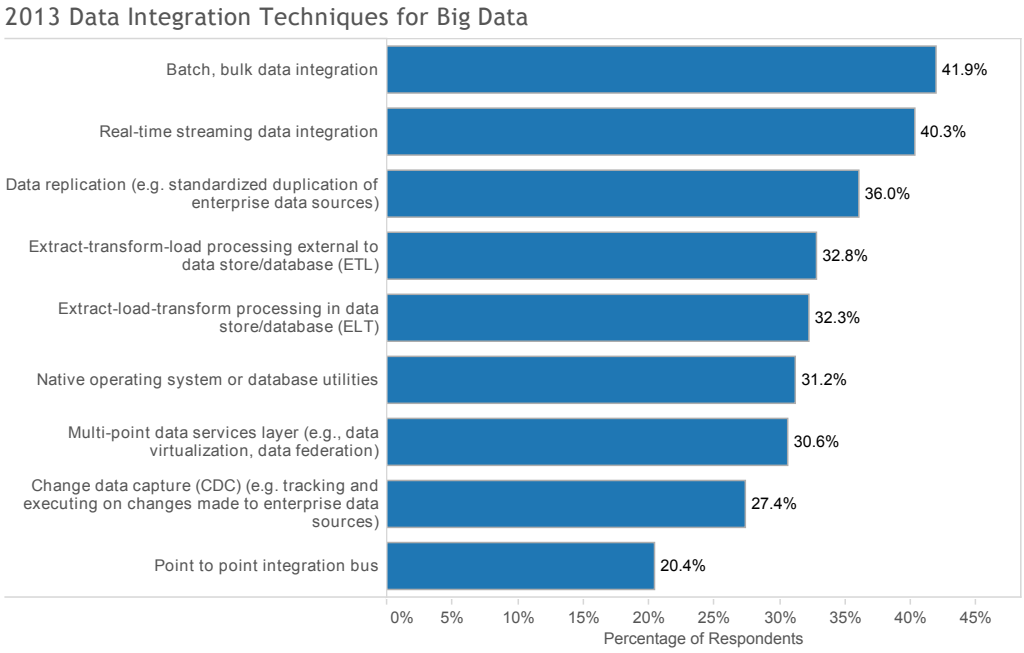


Figure 70

The next two indications are more consistent with relational-to-relational structure environments. Extract-transform-load processing external to data store/database (ETL) and Extract-load-transform processing in data store/database (ELT) techniques are most generally associated with moving structured data sets between platforms.

6.5.3. When to Apply Schema

In traditional data integration efforts associated with traditional operational or analytical environments, a structure, or data schema, is imposed on the data as it is being integrated. Often this is referred to as the application of "schema on write." In those traditional environments, it was important to match data structures in this fashion to ensure data management elements such as data quality and data governance. In addition, in those traditional environments, this task was relatively easy since a significant amount of the data was being moved from one relational structured environment to another.

However, in Big Data environments with the proliferation multi-structured datasets and variations in those structures, this task can take a significant amount of effort. For example, applying a **Schema on Write** to a JSON or XML dataset could result in:

- *Sparsely populated tables if the XML document format is relatively stable.*
- *Data from the source system being excluded if the target structure is not flexible.*
- *Constant adjustments to the target structure to accommodate the variations in JSON document format.*

In each of these situations, the concept of **Schema on Write** becomes difficult to implement or reduces the flexibility of the data integration process. Alternatives to **Schema on Write** are:

- ***Schema on Read:*** *This type of application of data structure is often referred to as "late binding schema" because the data structure is applied to the data late in the processing. This avoids the issue of making all data sources match a single schema upon load and it only applies the schema to the data associated with the processing and not ALL the data. Schema on Read requires increased processing power and coordination.*
- ***None:*** *This is a technique where no schema is applied to the data as it is integrated. Often this technique is used in online archiving situations where specific schemas have yet to be determined. Positively, this technique has a component of speed of data acquisition, but lacks the ability to truly organize the data since often there is a lack of visibility into the data structures.*
- ***Multi-schemas:*** *This technique is a combination of all the techniques above. Some datasets are processed on write. Some are processed on read. Some pre-processed datasets have their schema changed upon read. This technique shares the positive points of visibility to data structure and flexibility/speed of data acquisition. It also shares the downsides of the previous techniques.*

When asked how data models/schemas are applied to information in their Big Data environments, the top two answers of EMA/9sight respondents were associated with Multi-schemas and **Schema on Read**.

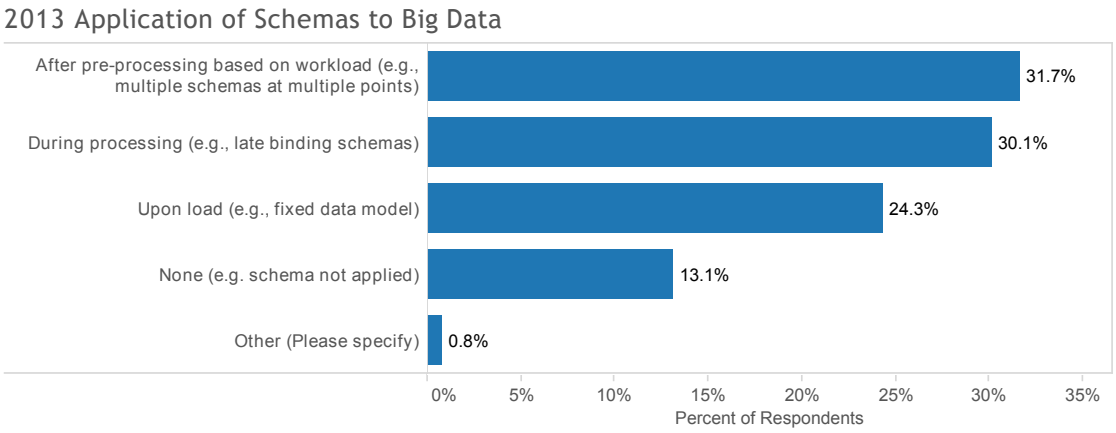


Figure 71

This indicates that EMA/9sight respondents are utilizing a more flexible approach than traditional **Schema on Write**. Yet as you can see from Figure 71, **Schema on Write** with fixed data models is still a significant part of Big Data strategies.

7. Case Studies

The following collection of case studies, provided by some of the sponsors of the 2013 EMA/9sight Big Data research, display how the trends of Big Data in 2013 are applied to impact the top and bottom line of the corporate balance sheet in business situations.

These cases studies show how organizations are tackling Big Data challenges across the following industries:

- Healthcare
- Social media and collaboration
- Mobile application distribution
- Database marketing services and technology
- Restaurant technology support
- Telecommunications

7.1. Brigham and Women's Hospital Handles Massive Data Volumes

Background

Brigham and Women's Hospital (BWH) is a 793-bed teaching affiliate of Harvard Medical School located in the Longwood Medical Area in Boston, MA. In addition to its biomedical research laboratories, BWH offers inpatient and outpatient services and clinics, neighborhood primary care health centers, and state-of-the art diagnostic and treatment technologies.

CASE STUDY INFORMATION

Customer name: Brigham and Women's Hospital

Customer domain: Healthcare

Vendor name: IBM

Product Area: Big Data Analytics, Data Management and Infrastructure provider

Opportunity

Traditional methods for analyzing large databases have become inadequate to deliver the potential that the research team at BWH envisioned for its growing trove of information. The institution was seeking an information management solution that would change the game, ultimately developing into a research tool that could learn over time, bringing the very latest drug effectiveness and interaction data right to patients' bedsides.

Solution

Drug risk awareness saves lives, which is why BWH recognizes the need to advance the complex analytics and computing power it deploys to help deliver "high-dimensional" pharmacoepidemiology research results. BWH implemented an information management solution that handles massive data volumes with ease and delivers analytics with unprecedented speed. The research team found it could immediately make full use of its large data sets and conduct multiple drug studies simultaneously. The solution is enabling researchers to design, test and apply brand new algorithms to help identify drug risk warning signals far more quickly. The institution intends to use the solution to automate a process for continuous drug safety monitoring and evolve the solution into a system that learns from prior results to improve its predictive accuracy.

Results

- Enables one of the department's novel algorithms, its high dimensional propensity scoring, to run 20 to 30 times faster than in its previous relational database environment.
- Enables the department to conduct basic analytic processing at two to three times previous speeds, with no change in code.
- Enables research studies on larger databases and exploration of previously inconceivable new research avenues.

7.2. Evernote Customer Experience Analytics

Setting

Evernote collects and analyzes large quantities of data about its users. Since 2008, more than 36 million people have created Evernote accounts, generating hundreds of terabytes of data, including 1.2 billion "notes" and 2 billion attachments. Notes can be text, Web pages, photos, voice memos, and so on – and they can be tagged, annotated, edited, sorted into folders, and manipulated in other ways.

CASE STUDY INFORMATION

Customer name: Evernote

Customer domain: Note taking and collaboration software

Vendor name: Actian

Product Area: Big Data Analytics and Cloud Integration software

Challenge

When this flood of data threatened to overwhelm Evernote's analytics system, the company modernized its analytic environment to handle big data – without breaking the budget. The provider of popular personal organization and productivity applications has moved from a conventional data warehouse to a modern hybrid of Hadoop and Actian's ParAccel Big Data Analytics Platform. To determine how to optimize the Evernote user experience, the company now analyzes 200 million events per day through a combination of Hadoop and the ParAccel Platform.

Benefits

- Evernote can store all their raw data in Hadoop and seamlessly load prepared data into ParAccel Platform for daily reporting as well as ad hoc analysis.
- Evernote answers many types of questions much faster than using Hive alone. It now takes three seconds to see which versions of Evernote Windows were most widely used in Germany during a particular week.
- Evernote has the ability to store all historical data, achieve fast ad hoc queries, and automate quality reports. Actian's ParAccel platform provides Evernote new daily insights into how customers use its products – and how those products can be continually improved.

7.3. Getjar Reduces Cost and Maintenance

Business Goal

Deploy a real-time Big Data collection and processing infrastructure to analyze mobile application download KPIs.

Challenge

Current in-house Hadoop instances were all batch-oriented and insufficient for realtime feed.

Solution

Use Treasure Data's real-time cloud data service for to solve business needs and reduce total cost of ownership.

Customer Reference

Simon Dong, Principal Architect, Getjar

CASE STUDY INFORMATION

Customer name: Getjar

Customer domain: Mobile application distribution

GETJAR

TREASURE DATA

Vendor name: Treasure Data

Product Area: Cloud Service for Big Data Acquisition, Storage and Analysis

From Batch Jobs in Hadoop to Low-Latency

"At Getjar, we discussed how we should address the increasing need for lower latency data feed as part of our data solution. To give you some context, Getjar has a lot of experience with Hadoop and big data. We've built several in-house solutions on top of Apache Hadoop technologies such as MapReduce, Pig and Hive for a variety of technical and business needs. However, up to that point, our in-house solutions were all batch-oriented; data were collected and processed once a day. For our upcoming lower latency business data feed, on the other hand, we had to collect data semi-realtime and run jobs every hour."

Cloud-based Solution Instead of On-Premises Hadoop

"One of the options was to implement a new in-house solution, but as Hadoop and big data practitioners, we knew the associated hardware, development and operation cost would be non-trivial.

That's why we decided to use Treasure Data instead of rolling out our own solution. Treasure Data's extensive features covered our needs, expedited the development greatly and gave us the confidence to move forward. We have been particularly happy with the very low total cost of ownership in terms of implementation and maintenance."

Getjar's Low-Latency Data Feeds

"We use td-agent to collect, convert and filter data and upload them to Treasure Data. Then, we use Treasure Data's REST API for hourly aggregation to populate our internal data warehouse (MySQL) from which our Tableau instances retrieve and visualize data."

Treasure Data Benefits 7.3.8. Stakeholders Across the Organization

"Treasure Data benefits all three stakeholders: Operations, Engineering and Business:

- Lowers the cost of implementing big data infrastructure while providing rich interfaces to allow flexible customization.
- Monitors and manages infrastructure.
- Enables lower latency KPIs and react more swiftly to changing business conditions."

7.4. Inferenda Customer Retention and Satisfaction Analytics

Setting

Inferenda is a market leader in database marketing services and technology, specializing in collating, aggregating and storing data. The company arms businesses with the ability to interact and analyze data as a commercial service. The company prides itself on its ability to support its customers, typically high-volume consumer marketing organizations or data resellers, through reliable self-service data count, fulfillment services and master data management. Their services help customers win more business and increase profitability.

CASE STUDY INFORMATION

Customer name: Inferenda

Customer domain: Database marketing services and technology

Vendor name: Actian

Product Area: Big Data Analytics and Cloud Integration software

Challenge

Inferenda's IT infrastructure was based originally on a Microsoft stack, but as their data volumes grew, their SQL Server database was unable to keep pace with their business and meet the needs and expectations of their customers. Inferenda and their customers often run up to 1,000 queries a day, many of which involve complex sorts and joins across multiple database tables totaling over a billion rows of data. When Inferenda started having several large clients extremely upset at having to wait 30 minutes or more for counts and orders information, they decided to switch to Actian's ParAccel Big Data Analytics Platform for blazing fast performance.

Implementing the ParAccel Platform required minimal tuning to meet and maintain Inferenda's performance needs. Since deploying the solution, Inferenda's self-service application for data count and order generation has translated into a measurable increase in customer retention and satisfaction. This has improved internal efficiencies as sales representatives no longer have to handle customer complaints, which in turn allows them to spend more time selling.

Benefits

- Higher levels of customer satisfaction because queries are running up to 60 times faster.
- Smaller infrastructure footprint because less disk space is needed.
- Less time responding to complaints and more time spent selling to new and existing customers.

7.5. Paytronix Integrates and Blends Big Data to Deliver Value to Customers

Challenges

Distributed Data Network

- Paytronix's loyalty and rewards programs software assists thousands of different restaurants. With terabytes of valuable restaurant and restaurant guest data in its system, Paytronix wanted to provide clients with deeper analysis capabilities to help them optimize their customer loyalty programs. The key factor here being the ability to integrate Paytronix's different types and sources of data.
- Extract, transform, load and store data more quickly and efficiently. With a new data warehouse, Paytronix could provide better, realtime analysis.

CASE STUDY INFORMATION

Customer name: Paytronix Systems

Customer domain: Technology for Restaurants

PAYTRONIX
Loyalty | Gift | Comp | Email

pentaho

Vendor name: Pentaho

Product Area: Business Intelligence and Analytical software

Improve Users' Relationship with Big Data

- With user-friendly dashboards and reporting tools, Paytronix could provide its customers with a clearer picture of guest behavior. To achieve more accurate snapshots of how Paytronix customers' restaurants were operating, they needed to improve existing tools.
- Paytronix needed a solution that would not require creating an entirely new software program, and not raise costs. They needed a synergistic solution that would integrate into their existing infrastructure, but was also cost effective.

Solutions

- Paytronix implemented the complete Pentaho Business Analytics platform, leveraging both the data integration and the business analytics suite.
- Pentaho Data Integration helped to facilitate Paytronix Data Insights, and provide ETL data from many different locations and sources.
- Pentaho Business Analytics identifies patterns that precipitate discounts, limited-time offers and visitors that will buy without an offer.
- Dashboards, Analyzer, Mondrian and InstaView are all used for data visualization to discover patterns and identify opportunities for merchants.
- Pentaho Concierge services created an OEM strategy that allowed Paytronix to embed Pentaho into Data Insights in less than two months.

Results

Paytronix rolled out a new Data Insights program to highlight three key improvements offered to their merchants. These highlighted the ability to dive deeper into guest data to identify actionable opportunities for driving visits and spending; the option of capturing data from a variety of sources beyond loyalty and gift programs including social media; and a visual interpretation of the data that enables end users to quickly uncover noteworthy trends.

Paytronix streamlined ETL capabilities provided an 80% reduction in data process time, improving both efficiency and saving money for its customers. Coupled with new reporting capabilities and

dashboards, Paytronix customers can explore new types of data previously unavailable due to the difficulty of data capture.

Restaurateurs using Paytronix's improved software can identify trends such as a steep increase in guest enrollment using mobile apps, for example, so that they can shift marketing priorities to quickly capitalize on opportunities. In addition, restaurants can identify poorly performing stores in time to adjust operational issues for optimal program performance. In turn, this has empowered Paytronix's customers and improved their competitiveness.

Summary

Paytronix Systems is the leading provider of gift, loyalty and email solutions for restaurants. With a portfolio of products serving over 200 restaurant chains and more than 8,000 locations, Paytronix is known as the restaurant industry's most innovative loyalty software provider. Paytronix wanted to provide clients with deeper analysis capabilities to help them optimize their customer loyalty programs. Paytronix needed a better way to store its data efficiently and Paytronix needed a simpler process to extract, transform and load its data. Their customers needed not only better real-time access to data, but also more user-friendly dashboards and reporting tools. With the complete Pentaho Business Analytics platform, Paytronix reduced its process time, optimized its data warehouse and provided an optimal end-user analytics experience. Now, Paytronix has bolstered its ability to provide its customers with a clearer picture of guest behavior. Today, their customers are empowered to leverage quick-to-access graphical interpretations of the data that has the greatest business impact – their guests' behavior.

> *"If you analyze data using an older set of cumbersome and time consuming tools, each question starts with ETL. Then you have to pull the data all the way through to get a look at the results. With Paytronix Data Insights, the Pentaho tools and our proprietary algorithms, customers experience an 80% reduction in ETL processing time, resulting in a self-service and cost-effective experience."* Andrew Robbins, President – Paytronix

7.6. Telecom Italia Anticipates Reducing Customer Churn and Responding to Service Issues

Background

Telecom Italia offers platforms and technological infrastructures that convert data and voice into advanced telecommunications services, plus the latest information and communications technology and media solutions. Along with its leading domestic position, the company has a major Latin American presence that generates around 34% of revenue.

CASE STUDY INFORMATION

Customer name: Telecom Italia

Customer domain: Telecommunications

Vendor name: IBM

Product Area: Big Data Analytics, Data Management and Infrastructure provider

Opportunity

In Italy, communications service providers face stiff competition. Customers demand higher levels of service and communications networks need to support increasing amounts of digital information. Despite collecting myriad data about its infrastructure, Telecom Italia could not proactively identify network infrastructure failure points or determine the root causes of service issues. As customers started to take their business elsewhere, the company was left with revenue gaps and high network operating and maintenance expenses.

Solution

After deploying a network intelligence solution, Telecom Italia gained a single, end-to-end view of its entire infrastructure: customer behavior, network performance, services levels and handset compatibility. The system integrates and correlates tens of billions of detailed transactional data with network, customer, device and other business data. Then, using sophisticated data analyses and Key Performance Indicators (KPIs), the company can proactively identify infrastructure and service failure points and address problems quickly. For example, if a KPI related to network performance is breached, the solution sends an alert and pinpoints the assets related to the event, allowing crews to take speedy remedial action before customers even know there is a problem or service is interrupted.

Results

- Boosts network performance insight by 100% by integrating multiple data sets in a single view.
- Anticipates reduced customer churn by improving customer service levels and minimizing network failure.
- Gains the ability to identify and respond to network issues proactively, before customers complain or drop service.

8. Methodology and Demographics

8.1. Research Methodology

EMA and 9sight Consulting crafted the Big Data user surveys that are the basis for this report. Before the survey was conducted, report sponsors were provided with a copy of the survey instrument. However, sponsors had no direct involvement in or influence on the survey creation, survey contents, survey execution, or any of the subsequent evaluation and analysis of the results for this report.

For this research, EMA and 9sight invited pre-qualified Business Intelligence (BI) and Information Technology (IT) professionals to complete an extensive web-based survey. These respondents were further qualified based on their responses to the following questions:

- What is your primary role in the usage and/or management of Big Data applications/technology within your organization?
- Which of the following best describes your company's primary industry?
- How would you describe the extent to which Big Data initiatives have been adopted within your business/organization?
- What is your relation to Big Data applications/products currently being used within your organization?
- At what phase of implementation are your business area /organization's Big Data initiative's project(s)?

Respondents who failed to qualify on these questions were rejected. As a result, all respondents (in addition to being independently pre-qualified through the initial invitation process) self-identified as being active participants with a working knowledge of current operational and analytical data management practices within their company that is presently researching, planning or implementing Big Data strategies.

8.2. 2013 Respondents

In 2013, 259 business and technology professionals responded to an invitation to provide their insights on Big Data strategies and implementation practices. To offer a balanced enterprise view of the subject, the respondent pool was also restricted. Business stakeholders represented 51% of respondents. Technologists were 48%. Professional services consultants in IT represented less than 1% of the response panel.

The 2013 survey instrument was executed between July and August 2013.

8.3. 2012 Respondents

For 2012, 255 business and technology professionals responded to the survey invitation. To provide balance, EMA/9sight restricted the respondent pool to an approximate mix of 45% business stakeholders, 45% IT participants and 10% IT consultants.

The 2012 survey instrument was conducted between July and August 2012.

9. Authors

Shawn Rogers
EMA

John Myers
EMA

Dr. Barry Devlin
9sight

9.1. About Enterprise Management Associates

Founded in 1996, Enterprise Management Associates (EMA) is a leading industry analyst firm that provides deep insight across the full spectrum of IT and data management technologies. EMA analysts leverage a unique combination of practical experience, insight into industry best practices, and in-depth knowledge of current and planned vendor solutions to help its clients achieve their goals. Learn more about EMA research, analysis, and consulting services for enterprise line of business users, IT professionals and IT vendors at http://www.enterprisemanagement.com or http://blogs.enterprisemanagement.com. You can also follow EMA on Twitter or Facebook.

9.2. About 9sight

Dr. Barry Devlin is founder and principal of 9sight Consulting (www.9sight.com). Barry is among the foremost authorities on business insight and one of the founders of data warehousing, having published the first architectural paper on the topic in 1988. With over 30 years of IT experience, including 20 years with IBM as a Distinguished Engineer, he is a widely respected analyst, consultant, lecturer and author of the seminal book, "Data Warehouse—from Architecture to Implementation" and numerous White Papers.

Dr. Devlin specializes in the human, organizational and IT implications of deep business insight solutions that combine operational, informational and collaborative environments. A regular contributor to BeyeNETWORK, Focus, SmartDataCollective and TDWI, Barry is based in Cape Town, South Africa and operates worldwide.

Corporate Headquarters:
1995 North 57th Court, Suite 120
Boulder, CO 80301
Phone: +1 303.543.9500
Fax: +1 303.543.7687
www.enterprisemanagement.com

2767.110713