CompTIA

# BIG DATA INSIGHTS AND OPPORTUNITIES

## FULL REPORT

RESEARCH


SECOND ANNUAL • SEPTEMBER 2013

www.comptia.org

## About this Research

CompTIA's *Big Data Insights and Opportunities* study builds on prior big data research conducted by CompTIA. The study focuses on:

- Tracking changes in how businesses capture, store, manage and analyze data
- Identifying key end user needs and challenges associated with data
- Gauging familiarity and incidence rates for big data initiatives
- Assessing IT channel partner perceptions of big data opportunities

The study consists of four sections, which can be viewed independently or together as chapters of a comprehensive report.

Section 1: Market Overview
Section 2: Bringing the Concept of Big Data into Focus
Section 3: Big Data and the Workforce Impact
Section 4: IT Channel Partner Perspectives of Big Data

This research was conducted in two parts.

**Part I: End User**
The data for this quantitative study was collected via an online survey conducted during June 2013. The sample consisted of 500 U.S. IT and business executives responsible for technical or strategic decisions affecting data at their company. Within the IT industry, this type of survey respondent is commonly referred to as an end user. CompTIA employed the services of a dedicated research panel provider to procure the sample. The margin of sampling error at the 95% confidence level for the results is +/- 4.5 percentage points. Sampling error is larger for subgroups of the data.

**Part II: Channel**
The data for this quantitative study was collected via an online survey conducted during April 2013. The sample consisted of 500 executives at U.S. IT firms, with most having some level involvement in the U.S. IT channel. The margin of sampling error at 95% confidence for aggregate results is +/- 4.5 percentage points. Sampling error is larger for subgroups of the data.

As with any survey, sampling error is only one source of possible error. While non-sampling error cannot be accurately calculated, precautionary steps were taken in all phases of the survey design, collection and processing of the data to minimize its influence.

CompTIA is responsible for all content contained in this series. Any questions regarding the study should be directed to CompTIA Market Research staff at research@comptia.org.

CompTIA is a member of the Marketing Research Association (MRA) and adheres to the MRA's Code of Market Research Ethics and Standards.

CompTIA

# BIG DATA INSIGHTS AND OPPORTUNITIES

## SECTION 1: MARKET OVERVIEW

RESEARCH


SECOND ANNUAL • SEPTEMBER 2013

www.comptia.org

## Key Points

- While already important, organizations of all types expect data to become even more critical to the success of their business. And yet, many companies acknowledge they need to do a much better job leveraging the data at their disposal. Nearly 8 in 10 executives agree or strongly agree with the statement "if we could harness all of our data, we would be a much stronger business." Because data growth shows no signs of slowing – IDC estimates the total volume of data doubles approximately every two years – this trend will only intensify.

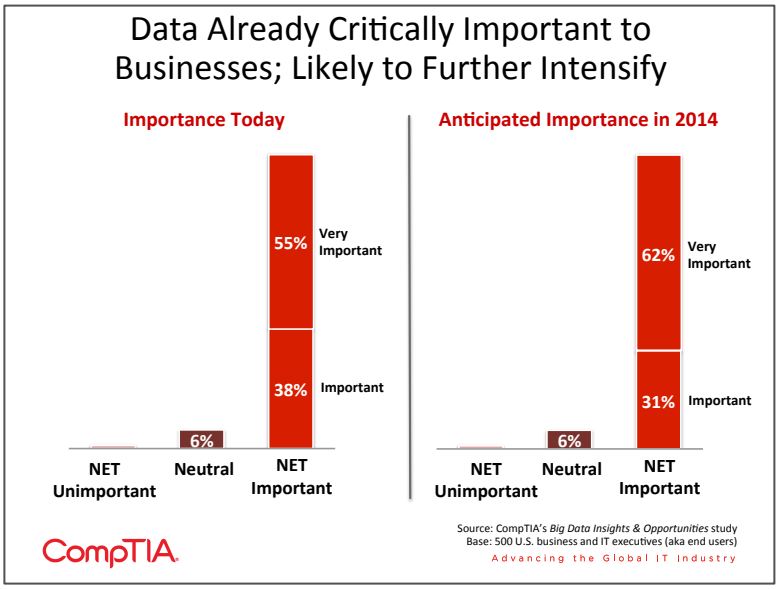
- The consequences of lagging behind in a data-driven world may become more pronounced. The participants in the CompTIA study cited lower productivity, lack of business agility, internal confusion over priorities and reduced margins due to operational inefficiencies as the top negative consequences of poor execution with managing and analyzing data. As companies move along the big data learning curve, many recognize the need for new skill sets, which will require investments in training and development to head-off worker shortages.

- Similar to any emerging technology category, sizing the big data market presents a number of challenges. While there are quite a few pure-play big data firms and a growing ecosystem of big data applications and tools, the challenge comes in separating the 'big data' component from the host of supporting and related technologies (think storage, data management or analytics).

## Overview

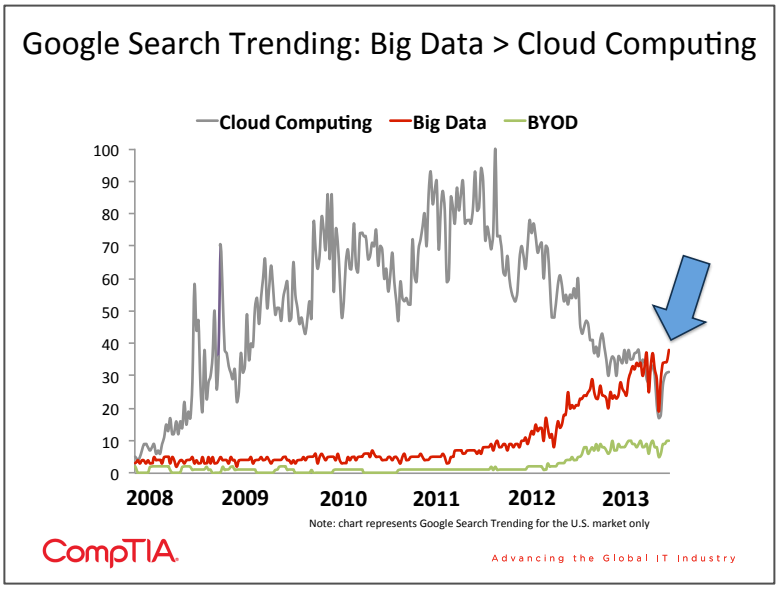
The total volume of worldwide data now doubles approximately every two years. Lower storage costs, more powerful distributed processing and new analytics tools have given businesses better ways to make sense of this ever-growing deluge of data. At the same time, businesses continue to wrestle with data silos, unstructured data, lack of real-time analysis and 'data noise' that interferes with decision-making.

While these are not new issues, nor is the importance of data a new phenomena, something has clearly changed.




Today, it is not uncommon to see data described as the 'currency' of the digital economy, suggesting a level of importance and value exponentially higher than in previous times. Getting to this point did not happen overnight, though, but rather through a series of developments on many fronts of the data story.

Taken together, these factors have ushered in the era of data. And, the aptly named concept of 'big data' emerges when data reaches extreme volumes, extreme velocities (the speed with which it is created or captured) and extreme varieties (different types of data).




While the threshold for what constitutes big data continues to evolve, businesses of all types will seek ways to unlock additional value from the data most relevant to them, be it on a large scale or a small scale.

Hype aside, the big data trend is still in its early stages and may eventually come to symbolize Amara's Law: the tendency to overestimate the effect of a technology trend in the short run and underestimate the effect in the long run.

## Sizing the Market

As with any emerging technology category, market sizing estimates vary. Different definitions, different methodologies and different time horizons can affect the output and interpretation. For example, as it relates to categorization, how much storage, data management or analytics revenue should be allocated to the big data market? Arguments could be made for various market sizing approaches.

With that caveat, the following estimates help provide context in understanding data-related growth trends.

- The research consultancy IDC estimates the worldwide volume of data will reach 3.6 zettabytes (1 billion terabytes) in 2013, and will keep growing at a rapid rate, doubling approximately every two years. By 2020, IDC projects a global data total of around 40 ZB.

- IDC predicts the global big data technology and services market will grow annually at 32% CAGR through 2016. The market is projected to reach $23.8 billion, up eightfold from $3.2 billion in 2010. Among this figure, $6 billion is expected in big data-related storage.
    - Note: the aggregate figure does not include data analytics software or services; IDC forecasts that segment of the market separately, projecting $71 billion in revenue by 2016.

- The research consultancy Gartner predicts big data will drive $34 billion of IT spending in 2013.

- The Wikibon big data community projects a market of $18.1 billion in 2013, up 61% year-over-year. Over the next four years the big data market is expected to approach $50 billion worldwide. Wikibon estimates a big data market distribution as follows:
    - 41% = hardware
    - 39% = services
    - 20% = software

- The research firm MarketsandMarkets predicts the worldwide Hadoop market will grow from $1.6 billion to nearly $14 billion over the next five years.

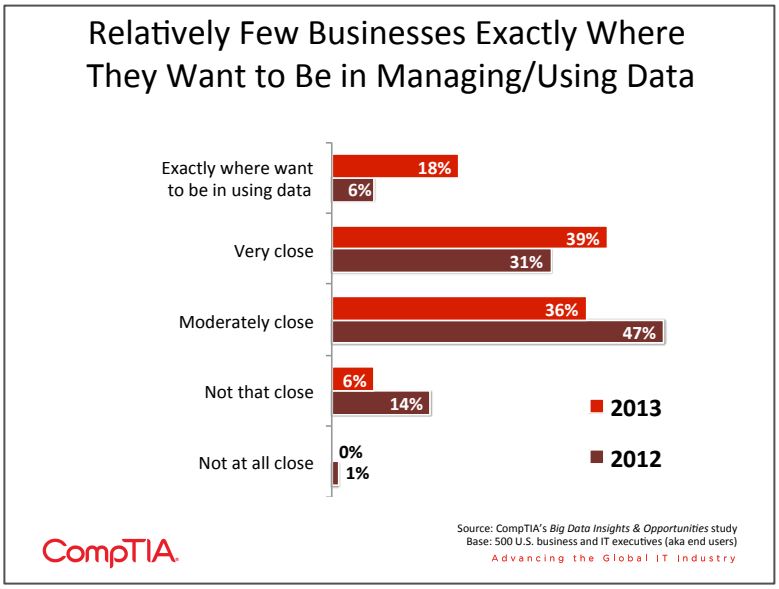
- According CB Insights, a consultancy that tracks investments, during the first half of 2013, U.S. big data startups raised $1.3 billion in funding across 127 deals. Over the past five years, big data investments attracted $4.9 billion in funding.
    - The firms attracting the most funding include: Cloudera, MuSigma, 10gen, GoodData and DataStax (Source: CrunchAnalytics)

- The range of vendors in the big data space is expansive. Storage firms, analytics firms, software firms, cloud firms and consulting and service firms make up the universe of players. A list of key firms with big data revenue estimates has been compiled by the Wikibon community and can be found here. Additionally, the CRN 2013 Big Data 100 vendor list can be found here.

## Assessing Data Challenges

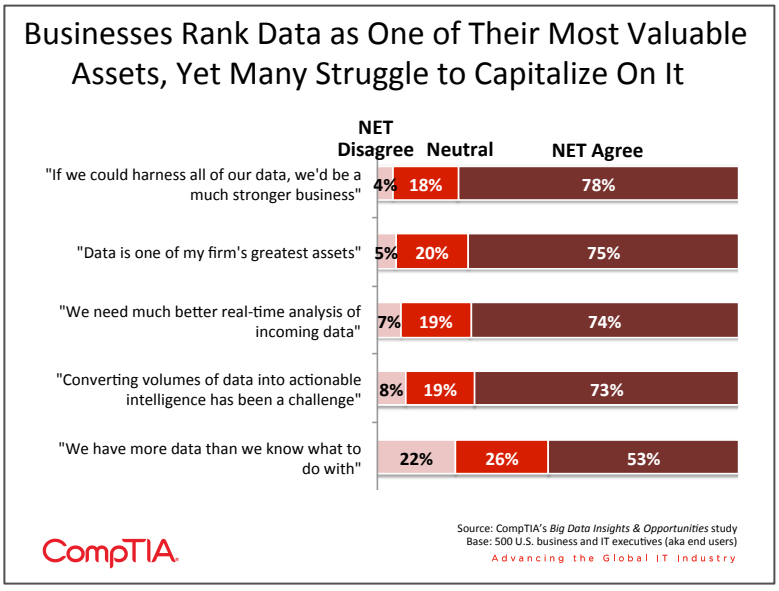
While most businesses recognize the growing importance of real-time access to actionable data, few have actually reached their data-related goals. Even fewer can claim to be anywhere near the point of engaging in a big data initiative.

According to CompTIA research, fewer than 1 in 5 businesses report being *exactly* where they want to be in managing and using data. Granted, this represents a high bar, but even when including those 'very close' to their target, it still leaves a majority of businesses with significant work to do on the data front.




Interestingly, the research suggests a notable jump in the number of firms reporting proficiency with data management and utilization, which may be a reflection of greater corporate focus on data management as a result of the big data trend. This sentiment seems to be corroborated by the 78% of survey respondents that report feeling more positive about big data as a business initiative this year compared to a year ago.

One of the strongest arguments for investing in data initiatives stems from the data point: nearly 8 in 10 executives agree or strongly agree to the statement *"if we could harness all of our data, we would be a much stronger business."*




A nearly equal number express a similar sentiment regarding the need for better real-time analysis and improvement in converting data into actionable intelligence.

See *Section 2* of this report for more details on the technical hurdles many organizations face in managing and analyzing their data.

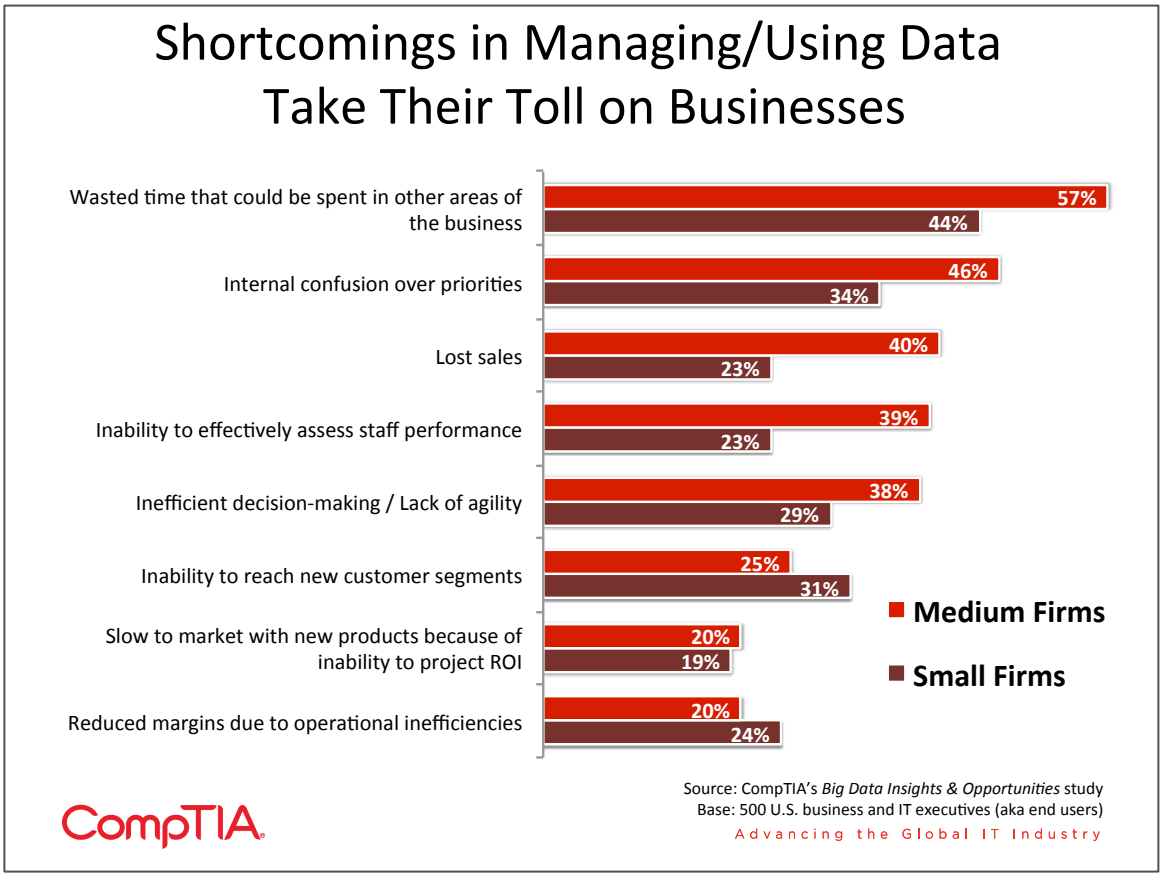
**Top 5 Costs Attributed to Shortcomings in Managing/Using Data**

1. Wasted time that could be spent in other areas of the business
2. Internal confusion over priorities
3. Inefficient or slow decision-making / Lack of agility
4. Inability to effectively assess staff performance
5. Lost sales and reduced margins due to operational inefficiencies

While some businesses may have made progress in an area of data management, many have not fully 'connected the dots' between developing and implementing a data strategy and its affect on other business objectives, such as improving staff productivity, or developing more effective ways to engage with customers (see table in Appendix).

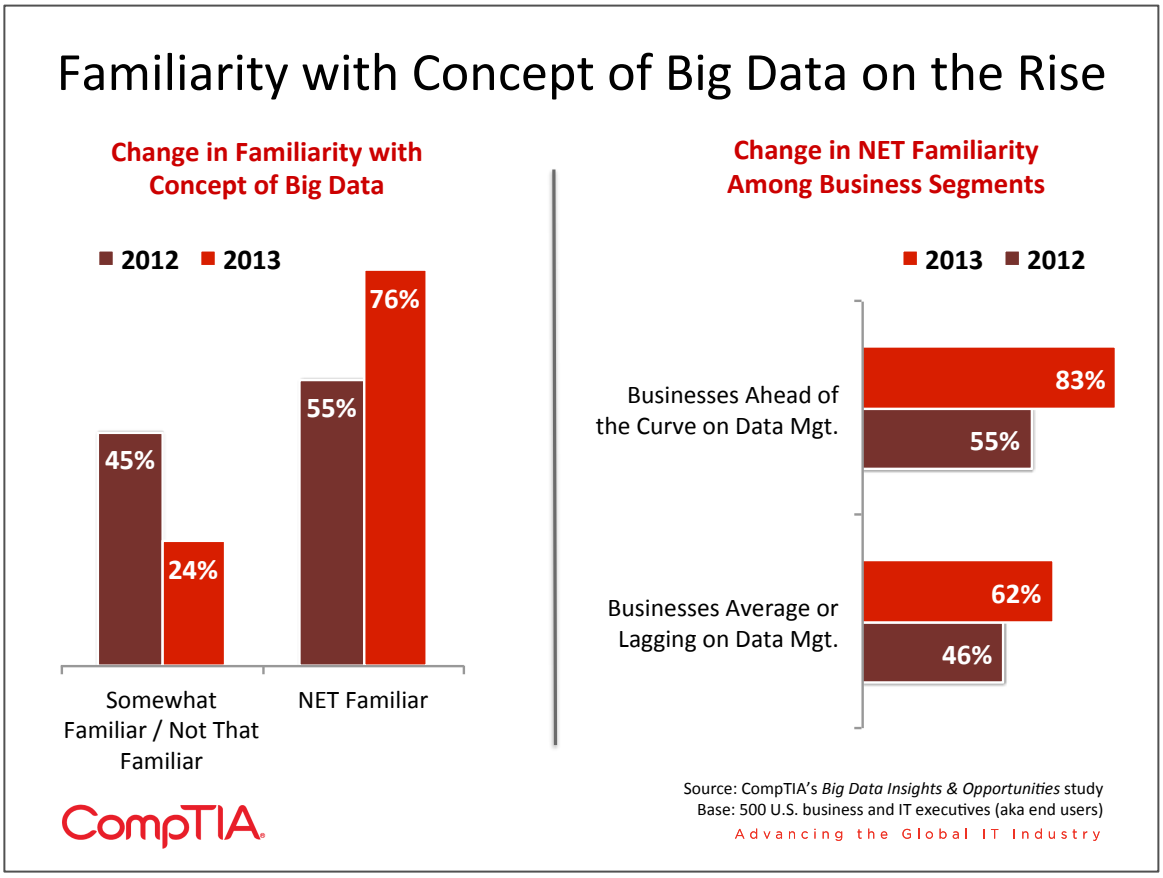
As businesses scale above 100 or so employees, complexities arise on a number of levels. Consequently, medium-size businesses voice more concern over shortcomings in the management and utilization of data. For example, 46% of medium-size businesses (100-500 employees) cite confusion over internal priorities as a consequence of poor data management compared to 34% of small firms (<100 employees).

For IT solution providers or vendors working in the big data space, this should serve as an important reminder to connect data-related solutions to business objectives, emphasizing outputs over the nuts and bolts of the inputs.



Using a basic definition of big data, CompTIA research found familiarity levels jumped twenty percentage points, from 55% to 76% year-over-year. Given the amount of media hype, this finding shouldn't come as a surprise. Keep in mind, though, familiarity rates are not meant to measure depth of knowledge of subject matter, but rather to assess top-of-mind awareness. Nonetheless, it serves as a useful proxy for interest levels and the degree to which customers are likely to be receptive to information and guidance on the topic of big data.

Among business executives that rate their firm as ahead-of-the-curve on data management, familiarity rates significantly higher than those lagging on data management. Hence, even with the jump in familiarity rates, there is still a sizable segment of the market on the bottom rungs of the learning curve.




While the CompTIA study did address the incidence of engagement in a big data initiative, the results should be used with caution. As noted previously, in an emerging market with evolving definitions and criteria, a big data initiative for one company may be a small data initiative for another. The consultancy Deloitte has used a threshold of 5 petabytes of data to be considered in the realm of big data, while IDC has used 100 terabytes; these criterion eliminate many data initiatives from big data consideration.

With that said, CompTIA found 42% of respondents claiming to be engaged in some of big data initiative. This translates to roughly double the number making the claim in 2012 (42% vs. 19%). Using either of the thresholds above suggests the 42% figure is high. This may stem from confusion or reflect the possibility of different users interpreting the concept of big data in different ways. As the market matures and businesses develop a better understanding of what is and what is not typically defined as big data, the accuracy of adoption figures will improve.

## Getting from Point A to Point B: The Workforce Factor

The big data trend sits firmly in the earliest stages of its life cycle. Moving along the adoption curve will require further advances on the technological front, as well as plenty of learning and knowledge gains to make it all work. The big data umbrella covers a wide range of skills, from deep technical to deep analytical and many combinations in between.

The McKinsey Global Institute calculates the demand for big data talent will far outstrip supply over the next few years. By 2018, the consultancy estimates a shortage of nearly 1.7 million workers in the U.S. alone. This includes a shortage of 140,000 to 190,000 workers with deep technical and analytical expertise, and a shortage of 1.5 million managers and analysts equipped to work with and use big data outputs. The obvious takeaway: many U.S. businesses will be unable to fully take advantage of big data initiatives because of an inability to find workers with the right skill sets and experience.

Gartner makes a similar claim. The firm predicts 1.9 million IT and analytical jobs will be created in the U.S. to support big data by 2015.

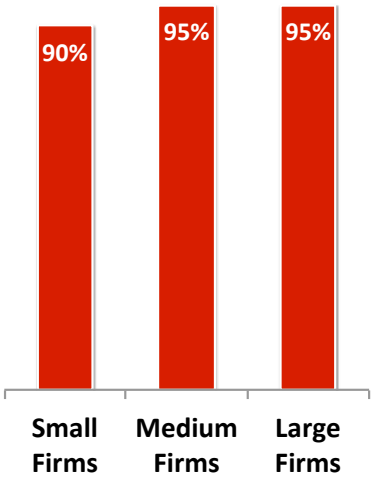
CompTIA's 2012 *State of IT Skills Gaps* study found companies currently place great importance on skills related to managing servers, data centers, storage and information. The emerging areas of analytics and big data-specific tools such as Hadoop are lower on the list, which makes sense given the relative newness of these technologies. It's likely future iterations of CompTIA skills gaps research will show gains in importance of big data-related skill sets.

See *Section 3* of this report for a deeper dive on data-related workforce issues.

## Appendix

# NET Importance of Data by Firm Size

**NET Importance of Data to the Business**




**Strategic Priorities for Improved Data Utilization Over Next 12 Months**

- **56%** Reducing costs / overhead
- **51%** Analyzing and improving internal workflows and communications
- **51%** Making better / faster business decisions
- **44%** Finding/implementing more effective strategies to reach new customers
- **42%** Further leveraging technology to improve business operations
- **33%** Better understanding industry's landscape, including competitive analysis
- **31%** More effectively bringing new products to market

Source: CompTIA's *Big Data Insights & Opportunities* study
Base: 500 U.S. business and IT executives (aka end users)

Advancing the Global IT Industry

# Businesses Seek Improvement in Many Areas of Data Analytics

| Data Analytics or Data Capability Type | Currently Doing Well | Doing, But Want to Improve | Want to Start Doing |
|---|---|---|---|
| Email marketing campaign effectiveness | 32% | 42% | 14% |
| Website traffic patterns | 28% | 47% | 16% |
| Real-time analysis of incoming data | 26% | 52% | 17% |
| Remote or mobile access to corporate data | 26% | 46% | 20% |
| Social media monitoring | 26% | 40% | 22% |
| Relationship analysis (e.g. X is highly correlated with Y) | 26% | 37% | 26% |
| Search capabilities across organization's many data sources | 24% | 50% | 17% |
| Customer profiling and segmentation analysis | 24% | 46% | 24% |
| Visualization capabilities (e.g. dashboards, etc.) | 24% | 46% | 19% |
| Pattern recognition | 24% | 37% | 26% |
| Metrics and Key Performance Indicators (KPIs) | 23% | 39% | 26% |
| Predictive analytics to forecast sales and other trends | 23% | 44% | 25% |

Source: CompTIA's *Big Data Insights & Opportunities* study
Base: 500 U.S. business and IT executives (aka end users)

Advancing the Global IT Industry

CompTIA

# BIG DATA INSIGHTS AND OPPORTUNITIES

## SECTION 2: BRINGING THE CONCEPT OF BIG DATA INTO FOCUS

RESEARCH


SECOND ANNUAL • SEPTEMBER 2013

www.comptia.org

## Key Points

- The three defining aspects of big data—volume, velocity, and variety—are really a description of the challenges businesses face in dealing with today's data. These aspects are being brought into focus as companies come to rely more heavily on data and as the tools become available for managing and analyzing data in new ways.

- One of the major challenges for businesses as they embark on new data initiatives entails having a good understanding of their current data situation. Eight out of ten firms report having some degree of data silos in their organization, with the number of firms reporting a high degree of silos rising from 16% in 2012 to 29% in 2013. Many emerging technologies associated with big data hope to tackle the challenge of disparate, unconnected data sources.

- Before considering the wide array of new technologies that are available for use today, companies must take a step back and assess the overall flow of their data and the objectives they hope to achieve. Within this flow, firms have a wide array of tools available, from traditional SQL databases, a growing set of NoSQL tools and the emerging distributed computing framework Hadoop.

## Bringing the Concept of Big Data into Focus

The starting point for understanding the underlying elements of big data begins with an examination of the big data definition.

> ***Big Data****: defined as a volume, velocity and variety of data that exceeds an organization's storage, compute or management capacity for accurate and timely decision-making (Source: The Meta Group).*

While variations exist, increasingly, this version is viewed as the consensus definition.

As noted in *Section 1*, it is important to remember that the challenges in dealing with data will be different from company to company, depending on the size of the organization, the skill in handling data, and the business needs for producing insights. This supports the concept of applying a situational or relative definition to big data. Consider this example:

An average SMB may manage 50 terabytes of data on a day-to-day basis, whereas Facebook manages 100 petabytes of user data (or, 2000 times more data). An initiative that creates 50 terabytes of new data would be a significant challenge for the SMB, though they would likely still employ more traditional data tools. Such an initiative for Facebook would be business as usual. The thresholds become relative as it relates to big data, as well as how the companies respond – the need to use emerging data management/analysis tools versus the viability of established tools.

### Businesses Contend with an Ever-Growing Volume of Data

| Data Type | Increasing Volume | No Change | Decreasing Volume | Don't Know/ NA |
|---|---|---|---|---|
| Email or IM | 68% | 29% | 2% | 0% |
| CRM-type data (e.g. customer records) | 63% | 34% | 3% | 1% |
| Log files (e.g. website, server, etc.) | 58% | 39% | 2% | 1% |
| Documents (e.g. Word, PowerPoint, Excel files) | 58% | 38% | 5% | 0% |
| Transactional (e.g. customer purchases) | 52% | 42% | 4% | 2% |
| Images or photos | 48% | 48% | 3% | 1% |
| Social or click streams (e.g. Twitter, web ad clicks, etc.) | 47% | 44% | 6% | 4% |
| Sensor data (e.g. RFID, weather, machine to machine) | 46% | 44% | 4% | 5% |
| Video | 43% | 45% | 8% | 4% |
| Geo-location data (e.g. maps, GPS locations, traffic, etc.) | 43% | 48% | 4% | 5% |
| Audio | 36% | 52% | 9% | 3% |

Source: CompTIA's *Big Data Insights & Opportunities* study
Base: 500 U.S. business and IT executives (aka end users)

CompTIA
Advancing the Global IT Industry

To put data volumes in perspective, an InformationWeek Analytics survey of U.S. businesses suggests 42% of companies have 100 terabytes of data or more, with 11% holding 1,000 TB or more. Undoubtedly some companies and some industries work with enormous volumes of data, but as a reality check, 6 in 10 businesses have a much smaller data footprint.

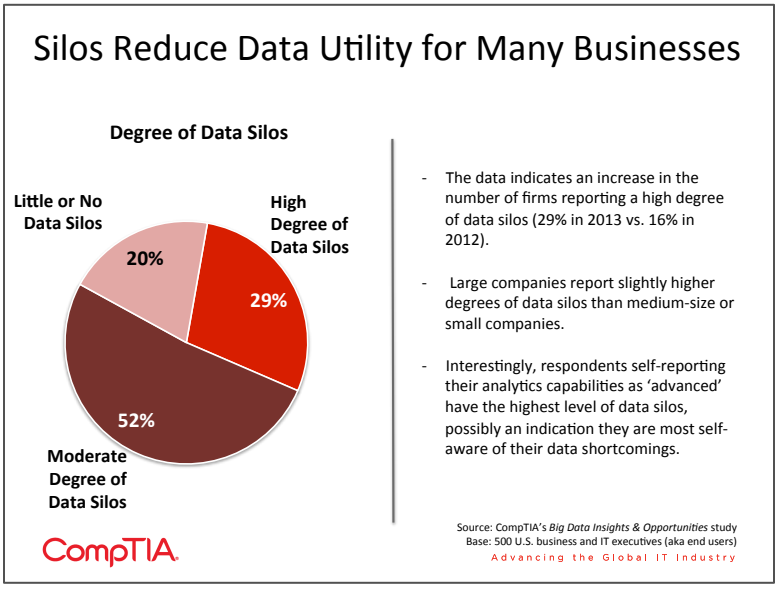
The velocity at which companies are able to consume data is mostly a function of the improvements in infrastructure, such as Wi-Fi networks or mobile broadband that allows data from mobile devices or sensor networks to be collected efficiently. Moreover, technology has enabled more businesses to assume the dual role of high volume, high velocity data consumer and data creator.

One of the ways in which velocity can be variable stems from situations involving major spikes in the influx of new data. Consider a department store during the holidays, attempting to understand buying patterns and react in real-time with pricing adjustments or staff messaging. Looking ahead to the era of the Internet of Things (IoT), whereby many more 'things' have an IP-based sensor, it's easy to imagine a scenario such as a natural disaster triggering a high-velocity burst of data that will need to be managed.

These examples point to a need for rapid scalability – a defining characteristic of cloud computing. For companies without sufficient infrastructure to handle highly dynamic workloads, cloud-based data management and analysis provides a practical platform to meet this type of need.

With IDC estimating that 80-90% of enterprise data is comprised of unstructured or semi-structured data (that is, data such as raw email, documents, images, videos, social streams etc. that cannot easily be put into rows and columns), variety may be the most long-standing data issue now addressable through new techniques. Companies can increasingly gain insights from these data stores that could not be placed into traditional relational database systems and accessed through SQL queries.

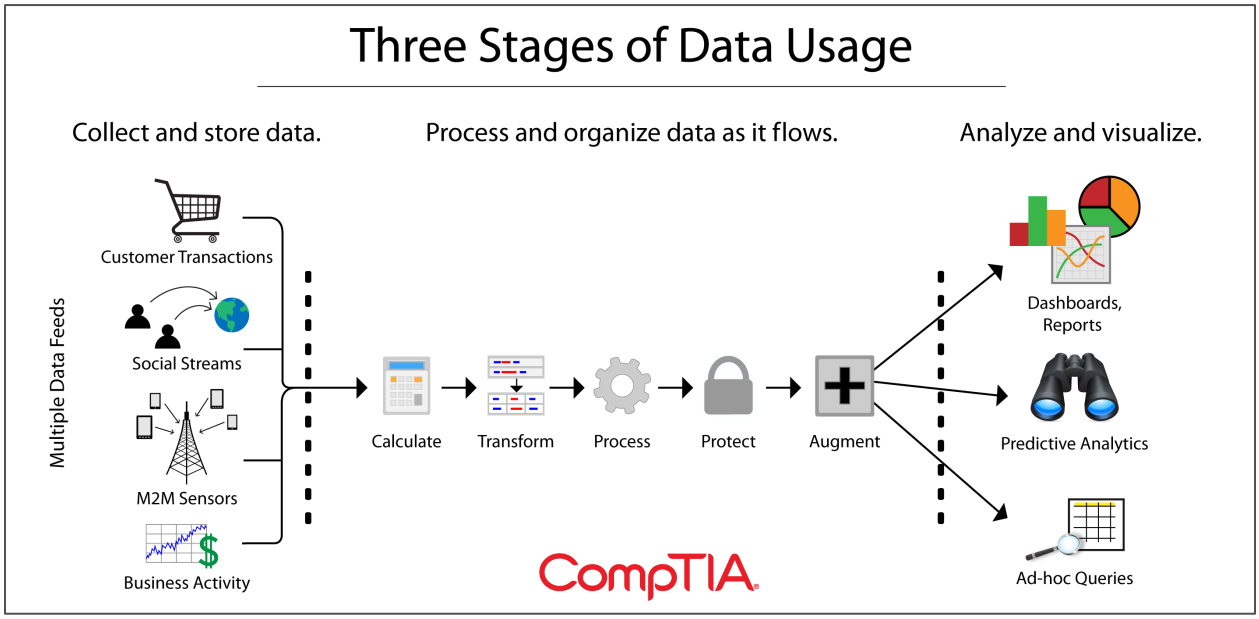
Since unstructured data necessitates a new approach, the challenge for most firms will be in building a solid foundation of data management practices on which to move forward into new territory. A prime example of an area in need of attention is data silos—collections of data that have grown within given departments but are not connected in a cohesive plan. As the volume of data continues to grow, these data silos can become larger problems preventing a company from fully leveraging its information. A data audit can identify where silos exist and provide a business with steps to take prior to adopting new data schemes.




Storage underpins any data initiative – see the Appendix for discussion on different approaches to storage.

## Behind the Scenes – A Look at Underlying Technologies

Before considering the wide array of new technologies that are available for use today, companies should step back and assess the overall flow of their data and the objectives they hope to achieve. In general, there are three stages to consider, with multiple possibilities in each stage that create a complex assortment of options.




With this understanding in place, companies may then turn their attention to an examination of the different technologies in the market. Since the early days of databases, sequential query language (SQL) has been vital to the storing and managing of data. Developed by IBM in the 1970s and popularized by Oracle as part of its commercially-available database systems, SQL has a long history and is well-known among developers for its ability to operate on centrally managed database schemas and indexed data. However, there are also limits to SQL. As data volumes grow, the architecture of SQL applications that operate on monolithic relational databases becomes untenable. Furthermore, the types of data being collected no longer fit into standard relational schemas. Two classes of data management solutions have cropped up to address these issues: NewSQL and NoSQL.

NewSQL allows developers to utilize the expertise they have built in SQL interfaces, but directly addresses scalability and performance concerns. SQL systems can be made faster by vertical scaling (adding computing resource to a single machine), but NewSQL systems are built to improve performance of the database itself or to take advantage of horizontal scaling (adding entire machines to form a pool of resources and allow for distributed content). Along with maintaining a connection to the SQL language, NewSQL solutions allow companies to process transactions that are ACID-compliant.

Although scalability and performance are addressed with NewSQL, there is still an issue with unstructured data that is not easily captured by relational databases. To handle unstructured data, many firms are turning to NoSQL solutions, which diverge further from traditional SQL offerings. Like NewSQL, NoSQL applications come in many flavors: document-oriented databases, key value stores, graph databases, and tabular stores. The foundation for many NoSQL applications is Hadoop. Hadoop is an open source framework that acts as a sort of platform for big data—a lower-level component that

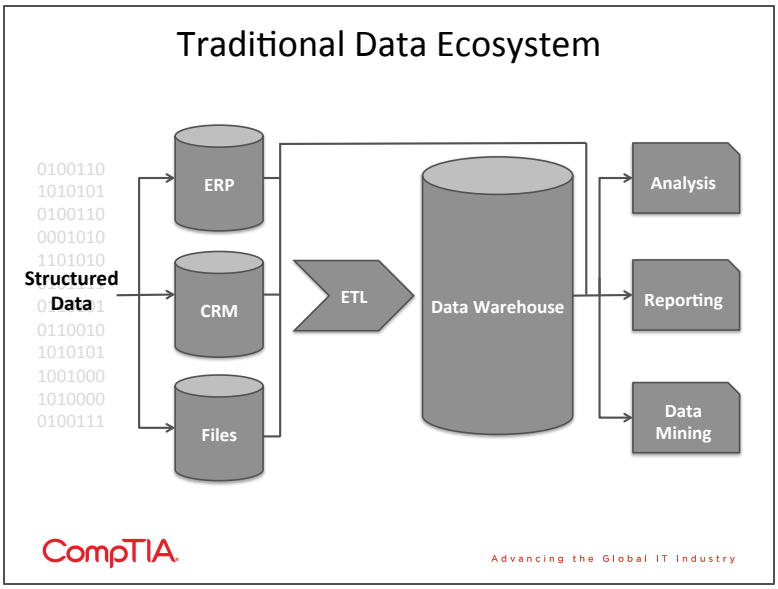
bridges hardware resources and end user applications. Hadoop is made up of two main pieces. The data is stored in the Hadoop Distributed File System (HDFS), which allows for storage of large amounts of data over a distributed pool of commodity resources. This removes the requirement of traditional relational databases to have all the data centrally located. The second component is the MapReduce engine, which utilizes parallel processing architecture to sort through and analyze massive amounts of data in extremely short timeframes. The ability of Hadoop to spread data and processing over a wide pool of resources creates the computing capacity to handle big data problems. For example, Twitter users generate 12 TB of data per day—more data than can be reliably written to a single hard drive for archival purposes. Twitter uses Hadoop to store data on clusters, then employs a variety of NoSQL solutions for tracking, search, and analytics.

| SQL | NoSQL |
|---|---|
| A programming language that forms the basis for the majority of relational database solutions on the market | A broad class of data management systems that addresses limitations of SQL relational databases |
| Data is stored in a single structure for consistency in operations | A distributed file system stores objects across a pool of commodity resources for high availability |
| Specific instructions are used to query and manipulate data based on the data being in a defined table | Different algorithms are used to query and manipulate data based on the type of solution (e.g. key-value, Big Table, document store) |

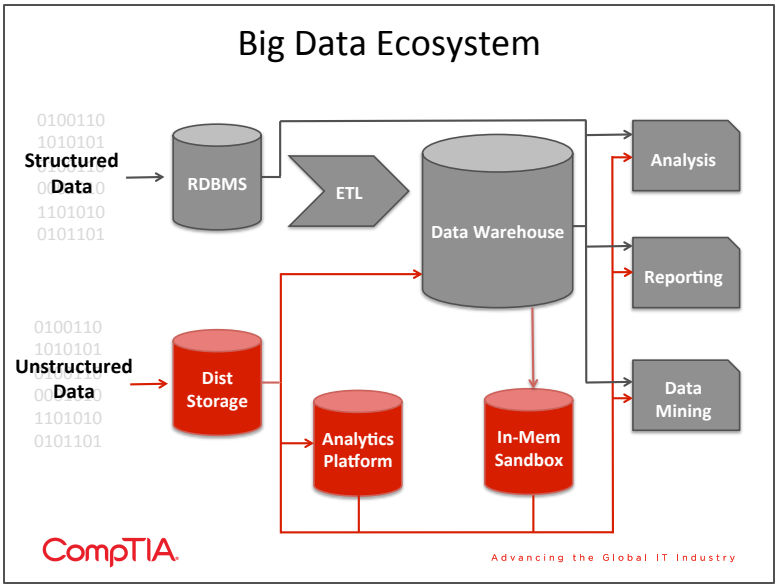
These NoSQL solutions are flooding the marketplace for multiple reasons. First, the nature of Hadoop as a platform means that it is not well suited for direct use by most end users. Creating jobs for the MapReduce engine is a specialized task, and applications built on the Hadoop platform can provide interfaces that are more familiar to a wide range of software engineers. Secondly, applications built on the Hadoop platform can more directly address specific data issues. Two of Twitter's tools are Pig (a high-level interface language that is more accessible to programmers) and HBase (a distributed column-based data store running on HDFS). Another popular tool is Hive, a data warehouse tool providing summarization and ad hoc querying via a language similar to SQL, which allows SQL experts to come up to speed quickly.

Although Hadoop is able to speed up certain data operations related to storage or processing, the initial release was not well suited for real-time analysis since it was designed to run batch jobs at regular intervals. The current stable release (1.2.1, released in August 2013) still follows this framework, but an alpha release (2.0.5, release in June 2013) features an upgrade of the MapReduce framework to Apache YARN. According to the release manager, Arun Murthy, this helps move Hadoop away from 'one-at-time' batch-oriented processing to running multiple data analysis queries or processes at once. Other tools, such as the Cassandra database and Amazon Dynamo storage system or Google's BigQuery, replicate some of Hadoop's features but seek to improve the real-time analysis capability, as well as ease of use.

High-level views of data ecosystems show the way that big data elements complement traditional elements and also add complexity. While some of the transitions look similar, such as a move from individual data silos into a data warehouse, the transactions are different—migrating unstructured data is a different operation from the standard Extract, Transform, Load (ETL) processes that are well established. In addition, there are new possibilities available to those analyzing data, such as in-memory databases using SSD technology.




Given the distributed nature of big data systems, cloud solutions are appealing options for companies with big data needs, especially those companies who perceive a new opportunity in the space but do not own a large amount of infrastructure. However, cloud solutions also introduce a number of new variables, many of which are out of the customer's control. Users must ensure that the extended network can deliver timely results, and companies with many locations will want to consider duplicating data so that it resides close to the points where it is needed. The use of cloud computing itself can create tremendous amounts of log data, and products such as Storm from the machine data analyst firm Splunk seek to address this joining of fields.

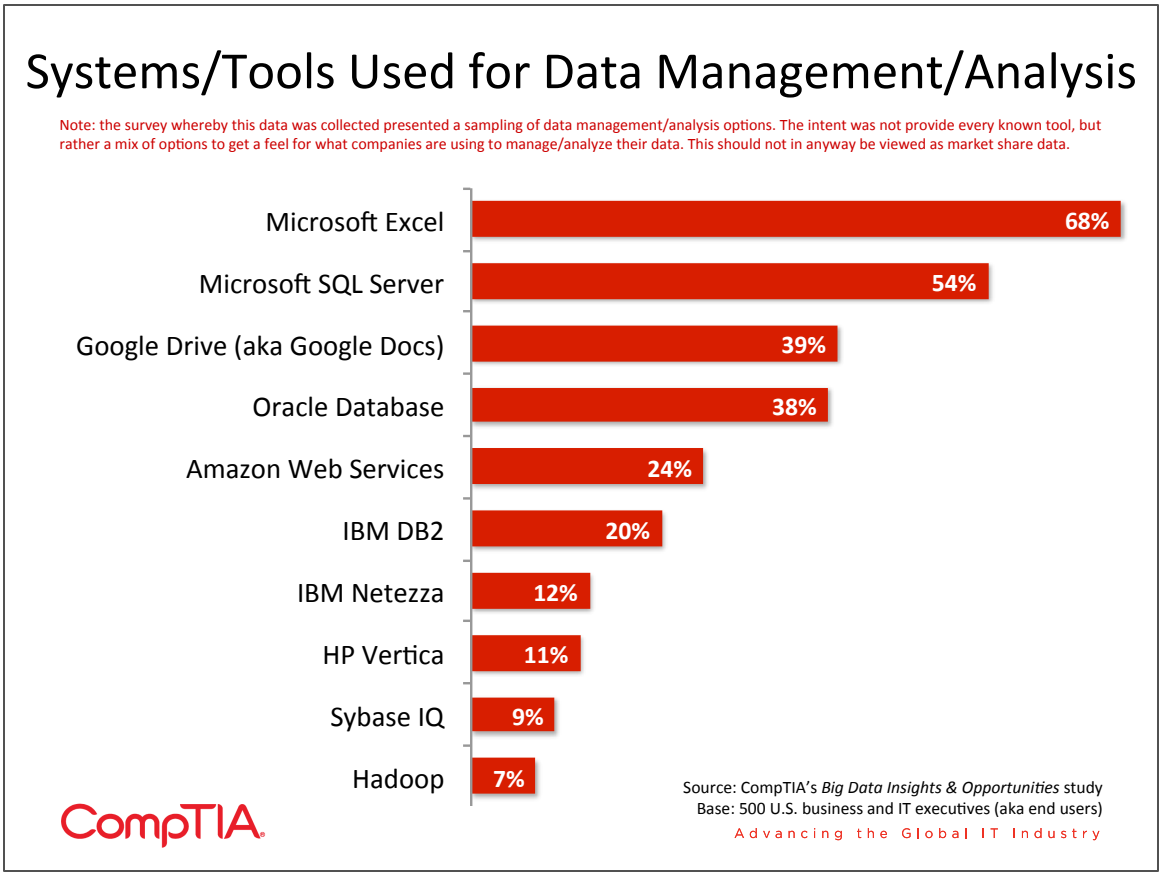



Companies on the leading edge of the big data movement, then, are finding that there is no one-size-fits-all solution for storage and analytics. Standard SQL remains sufficient for certain operations, and both NewSQL and NoSQL solutions have specialized benefits that make them worthwhile. As products continue to evolve (such as NewSQL systems handling unstructured data or NoSQL systems supporting ACID), the classifications may fade in importance as specific solutions match up with specific problems. With data streams constantly flowing from various sources, there are opportunities to improve services or trim costs by rapidly digesting and analyzing the incoming data. As products evolve and are introduced to meet needs, there will also be a growing need for education, product selection, and services surrounding big data initiatives.

## Appendix A

With only 16% of firms rating themselves at an advanced level of data analysis, it is not surprising to see that the most popular data tools tend to be the simpler ones. The survey results reveal some interesting usage patterns among companies of varying size and capability.

- While Microsoft Excel is used fairly consistently across small, medium, and large businesses, Google Drive has noticeably more uptake among small businesses (35%) than medium (31%) or large (39%). Installation and support for Google's tool suite is becoming a necessary skill to have alongside support for Microsoft's more established programs.
- Use of Amazon Web Services is surprisingly consistent across small, medium, and large businesses. This would indicate that most firms are viewing AWS as a management tool, and have not yet explored the capabilities built in for analysis. This is an area of opportunity for companies that do not have the capabilities to build their own analysis framework.
- For all the focus on Hadoop, only a small segment of companies are currently using the system, including only 11% of firms that rated themselves at an advanced level of data analysis. Even the 7% overall adoption may be somewhat inflated—5% of small businesses reported using Hadoop, and even that small percentage seems high for companies with fewer than 100 employees and less technical capability.



## Appendix B

In 2010, Google's Eric Schmidt claimed that the amount of data generated globally every two days matched the amount generated from the dawn of civilization up through 2003 (approximately 5 exabytes). Similar to the IDC prediction, IBM projects 35 zettabytes (35 billion terabytes) of data will be generated annually by 2020. All that data needs to be stored somewhere, and with many solution providers having existing knowledge of storage solutions, this is a good entry opportunity into data issues.

CompTIA's inaugural Big Data study more deeply explored storage usage trends. While this report does not contain the same level of detail, solution providers should still be familiar with some of the main forms of storage, especially since many of these may be combined to create an overall solution.

- **Rack Storage:** The most popular form of storage in the 2012 report, the implementation of discrete physical storage systems is the simplest in concept. Also known as Directly Attached Storage (DAS), it simply connects a storage solution to a given server. However, it is also the least flexible and can potentially be inefficient, especially when scaling out.
- **Virtual Storage:** As companies virtualize their workloads, the maximum benefit is gained when storage is also virtualized. This can be a major challenge, though: a May 2013 study conducted by DataCore Software found that 44% of companies cite storage costs as "somewhat of an obstacle" or a "serious obstacle" preventing them from virtualizing further. There are two main types of virtualized strorage:
    - **NAS (Network-Attached Storage):** A single storage device that can be allocated among several servers (either physical or virtual). A NAS operates at the file level and typically uses TCP/IP over Ethernet, so it is generally easier to manage. Servers see NAS devices as network storage.
    - **SAN (Storage Area Network):** A device or network of devices connected through iSCSI or Fibre Channel and operating at the block level. Thanks to this architecture, servers see SAN storage as local storage, allowing for faster, more reliable operation. SANs are significantly more expensive and require more specialized skill to implement.
- **Cloud Storage:** A component of Infrastructure as a Service (IaaS) offerings that give companies storage options that reside in a cloud provider. Some of these storage options are related to virtual instances that are created, and companies can choose whether the storage will persist or disappear after a virtual instance is deleted (similar to DAS). Other options are purely storage, and can be used in the way a company would use a NAS or SAN or be used for backup and redundancy.
- **Solid State Devices (SSD, aka flash):** As the desire for rapid data processing grows, SSDs are becoming more popular. Although cost is falling, SSDs are still significantly more expensive and do not provide the top capacity in a single unit. Many architectures may use SSD in select situations, with traditional storage still comprising the bulk of the solution. In a cloud offering, SSDs are likely in play wherever high I/O operations are advertised. SSD arrays are increasingly being used in conjunction with DRAM (dynamic random access memory) to create in-memory systems such as SAP's HANA database. These systems can process huge amounts of data in a fraction of the time compared to traditional systems, appealing to those companies with real-time analysis needs.

Regardless of the storage technology, without robust backup and recovery, companies put their operations at risk. See CompTIA's research brief on *Business Continuity/Data Recovery (BC/DR).*

CompTIA

# BIG DATA INSIGHTS AND OPPORTUNITIES

## SECTION 3: BIG DATA AND THE WORKFORCE IMPACT

RESEARCH


SECOND ANNUAL • SEPTEMBER 2013

www.comptia.org

## Key Points

- Moving along the adoption curve for big data will require further advances on the technological front, as well as further developments in experience and expertise among users. The big data umbrella covers a wide range of skill sets, from deep technical to deep analytical and many combinations in between. The McKinsey Global Institute calculates the demand for big data talent will far outstrip supply over the next few years.

- CompTIA's 2012 *State of IT Skills Gaps* study found IT and business executives rate skills associated with server, data and storage management as the highest data-related skills priorities. Data analytics and big data skills rate slightly lower, which given where they are in their life cycles is not surprising. Expect this to change over time, though.

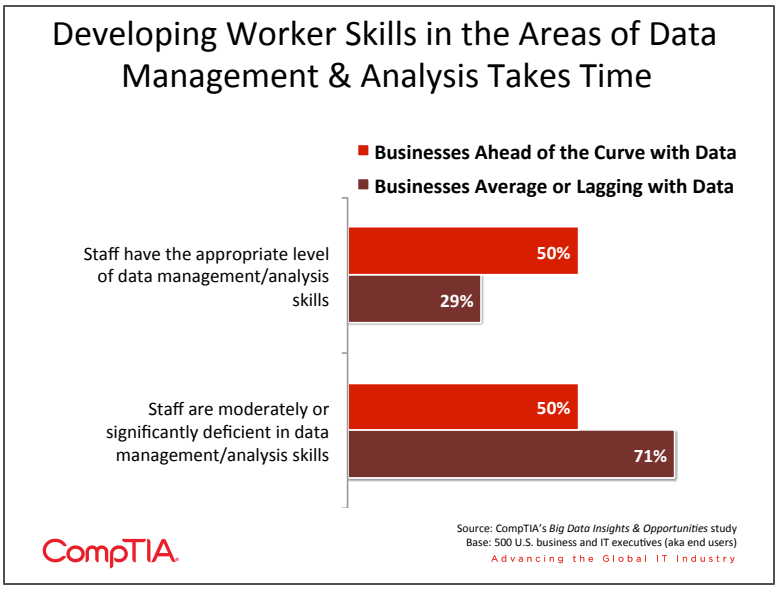
- To satisfy data management and analytical skills needs, businesses plan to take a number of steps. Many will employ a hybrid approach of leveraging in-house staff, supplemented with assistance from outside IT solution providers or other experts. Sixty-six percent of firms plan to invest in training for current employees, while 43% signal an intent to hire new workers with data-specific expertise.

## As the Big Data Trend Gains Momentum, Workforce Plays Catch-up

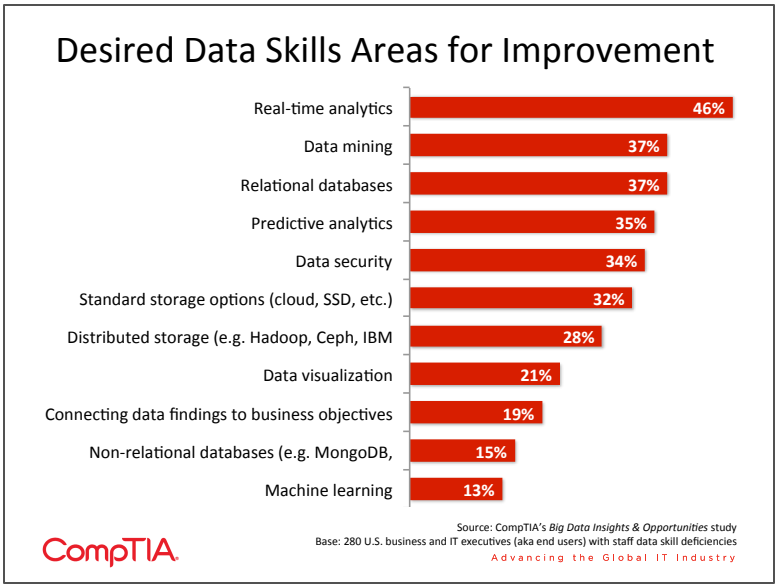
Few investments in technology succeed in delivering maximum benefits without a corresponding investment in human capital. In the realm of big data, this relationship is especially critical.

To recap from *Section 1*, the majority of businesses acknowledge a need for improvement in managing and using data. Consistent with this sentiment, 6 in 10 report a corresponding need to boost employee skill levels on the technical or business side of data management and analysis.

As expected, businesses that have already achieved a degree of sophistication with data are better positioned on the skills front, but even half of this segment indicates some level of skills deficiencies. At the other end of the spectrum, an overwhelming percentage of the data laggards (7 in 10) report skills gaps.




Interestingly, survey respondents with an IT-centric role appear more likely to believe their organization is sufficiently skilled with data management and analytics compared to their business-centric colleagues (44% vs. 29%). This may reflect a situation common to many organizations: an adequate infrastructure for aggregating and managing data, but insufficient capabilities in converting raw data into real-time actionable intelligence. Consequently, business executives may find more fault with the state of skills under this scenario than infrastructure-focused IT staff.




Realistically, most businesses will tackle skills deficiency challenges through a series of fits and starts. Because data initiatives often involve many moving parts, organizations will inevitably be forced to contend with the proverbial "weakest link" in the data chain before achieving optimization. While it may be premature to establish a precise checklist of must-have data skills, the research does provide a few clues as to where training, education and certification investment dollars may flow.

To satisfy their need for data management and analytical skills, businesses plan to take several steps:

**66%** Provide training to existing employees in the areas of data management/analysis
**43%** Hire new employees with expertise in data management/analysis
**35%** Utilize current assets, working to improve along the way
**30%** Contract with outside consultants or vendors that specialize in IT
**28%** Contract with outside consultants or vendors that specialize in business/management
**21%** Contract with outside consultants or vendors that specialize in digital strategy
**16%** Contract with outside consultants or vendors that specialize in marketing/customer insight

As expected, many will employ a hybrid approach involving in-house staff, supplemented with assistance from outside IT solution providers or other experts, and for some companies, the hiring of new employees.

Sixty-six percent of respondents plan to invest in training for current employees, an increase of 13 percentage points over the prior year's rate. Training may include vendor-based training (see chart on previous page for examples), training/educational tracks at conferences, training from providers such as Global Knowledge, or possibly even courses from a community college or 4-year university.

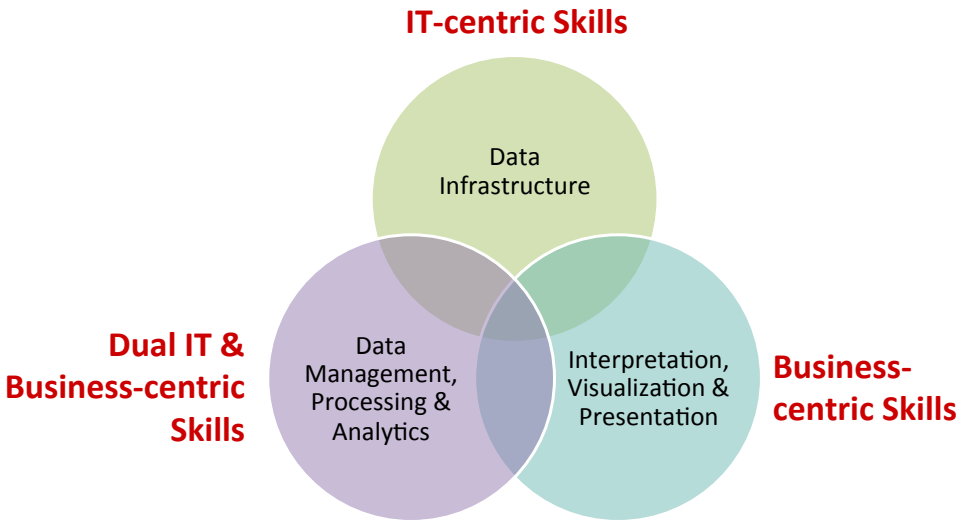
Forty-three percent of businesses signal an intent to hire new staff with data management or analytical expertise, an increase over the 2012 hiring intent rate. Even when factoring in the steady increase in data initiatives, it must be acknowledged that hiring intent is sometimes overstated. The executives responding to the survey may have every intent of adding headcount, but unexpected factors could derail plans – business conditions change, business priorities change, challenges emerge in finding the right candidate and so on. Nonetheless, even if slightly fewer businesses than 4 in 10 hire new staff, it's still a significant increase in demand for data savvy workers.

As businesses evaluate options for tapping outside expertise, several factors may influence their selections. Traditional management consulting firms (think Accenture, McKinsey & Company, Booz Allen, etc.) have always offered data analysis and related services. Now, many of these firms are working to capitalize on the big data trend by expanding their offerings to other areas of the organization, including the IT department. Similarly, traditional technology vendors and solution providers seek to move beyond the back-office and directly engage with front-office personnel responsible in a capacity more likely to impact the bottom line. Lastly, firms that may not have traditionally been viewed as having a role in data initiatives are also working to carve out a niche. Consider a digital marketing and PR firm that manages online advertising campaigns, social media and related digital strategies. These offerings all involve data and each can be amplified with the right data management and analysis tools, so it's not a stretch to think of these types of firms as moving into the big data space.

With data initiatives being driven by executives across the organization (see chart on subsequent page), expect personal preferences to come into play. For example, a CFO may have a natural proclivity to engage with a management consulting firm, a CMO with a digital marketing firm, a CIO with a technology vendor and so on.

Outside consultants and solution providers best able to navigate the technical and the business side of big data will be best positioned to meet customer needs. Going a step further, those with industry sector specialization (e.g. retail), or functional area specialization (e.g. HR) will be even better positioned.

# Big Data Skills Span Many Functional Areas




CompTIA

Advancing the Global IT Industry

# Data-Related Workforce Framework

| | Examples of Roles* | Examples of Certifications* |
|---|---|---|
| **Data Interpretation & Visualization** | – Data analytics / BI<br>– Data scientist<br>– Presentation / Visualization | – IBM Certified Specialist – Netezza<br>– MicroStrategy Certified Developer<br>– SAS Certified Predictive Modeler<br>– Tableau Certified Professional |
| **Data Management & Processing** | – Database administrator<br>– Architecture / developer<br>– Data / application integration<br>– Data lifecycle management | – Microsoft Certified DBA<br>– Oracle DBA Certified Master<br>– Cloudera Certified Administrator for Apache Hadoop (CDH3) |
| **Data Infrastructure** | – Storage<br>– Data center<br>– BC/DR<br>– Security<br>– Data capture | – Cisco CCNP Data Center<br>– EMC Storage Administrator<br>– CompTIA Storage+ powered by SNIA<br>– VMware Certified Professional: Datacenter Administration |

CompTIA

*Examples are for illustrative purposes ; not meant to be an all inclusive list or not meant to represent market shares.

## Average Salaries for Data-Related Positions

| Job Role | 2012 Mean Salary | Yr/Yr Change |
|---|---|---|
| Data architect | $114,380 | +5.0% |
| Database administrator | $94,430 | +2.9% |
| Business analyst | $88,887 | +3.4% |
| Business intelligence specialist | $101,854 | +2.1% |
| Data warehouse specialist | $101,061 | +6.1% |
| Hadoop specialist | $115,062 | NA |
| NoSQL specialist | $113,031 | NA |

Source: Dice, a leading career site for technology professionals

In addition to the positions above, the emerging role of data scientist will increasingly be viewed as a necessary component to any big data initiative.

However, similar to the trend itself, the data scientist job description continues to evolve. One of the world's largest aggregators of data, Facebook, posted an opening for a data scientist last year, providing insights into what this type of position may entail.

**Sample of key responsibilities for Facebook data scientist job opening:**

- Answer product questions by using appropriate statistical techniques on available data
- Drive the collection of new data and the refinement of existing data sources
- Develop best practices for instrumentation and experimentation and communicate those to product engineering teams
- Communicate findings to product managers and engineers

**Sample of skill requirements for Facebook data scientist job opening:**

- Comfort manipulating and analyzing complex, high-volume, high-dimensionality data from varying sources
- A strong passion for empirical research and for answering hard questions with data
- Experience working with large data sets, experience working with distributed computing tools a plus (Map/Reduce, Hadoop, Hive, etc.)
- Ability to communicate complex quantitative analysis in a clear, precise, and actionable manner

As expected the position description is heavy on quantitative analytics and technical skills, which is probably common to most data scientist roles. It's the other areas of the job that likely vary from firm to firm. In Facebook's case, the description indicates the data scientist must do more than produce interesting output. He or she must be able to positively impact the bottom line. The emphasis on communication with product managers, actionable guidance and the connection to product questions makes the position "real world" in nature, rather than purely theoretical work.

## The Catalyst for Data Initiatives

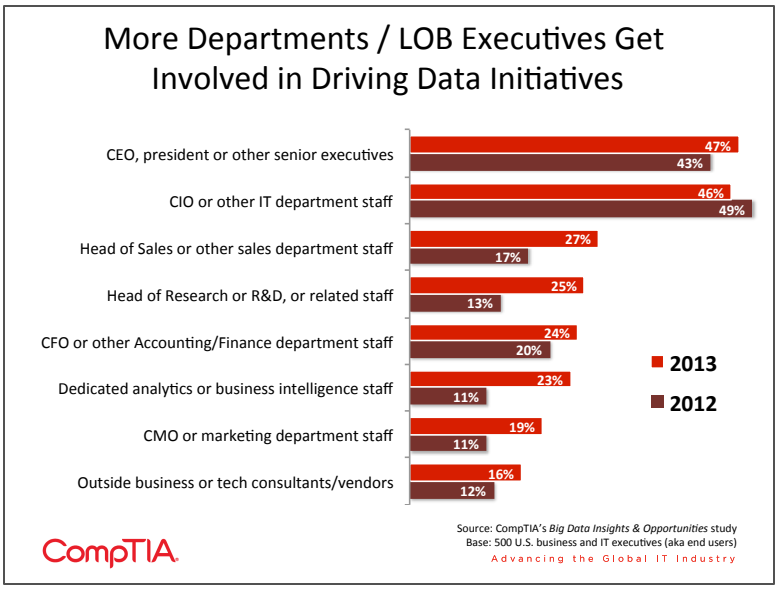
The concept of the 'democratization of IT' has been around for some time. Thanks to developments such as cloud computing, social networks and BYOD, employees across an organization are increasingly empowered to deploy technology as they see fit. For better and for worse, the transition from centralized control (aka IT department/CIO control) to decentralized decision-making along business unit lines is well underway. Past CompTIA research has consistently shown an uptick in technology purchases, both authorized and unauthorized, taking place outside the IT department. As further confirmation, over the next few years, Gartner predicts line-of-business (LOB) executives will be involved in 80% of new IT investments.

The big data trend will further accelerate the democratization of IT. Because data now permeates every functional area of an organization, data-related initiatives may originate from any number of departments or line-of-business (LOB) executives. As seen in the chart below, a diverse range of business executives have a hand in driving data-related initiatives. Starting with the CEO and then on down the chain, the end-users of data – be it financial, operational or customer data – typically have the strongest incentive to seek new approaches to leveraging data.

From a workforce perspective, it presents an interesting question: how best to develop staff equipped to drive big data initiatives. Does it make more sense to cross-train IT-centric staff on the business intelligence/analytics/interpretation side of big data, or cross-train business-centric staff on the technical side of data management and utilization? Using the Venn diagram above, the ideal position incorporates skills across each functional area. While companies may strive to recruit or develop workers with deep subject matter expertise across various knowledge areas, it will take time for the big data labor force to mature.

Consequently, opportunities will exist for third party firms, such as IT solution providers, that can lend expertise in one or more areas of data processing, management or analysis.

Additionally, because of the diverse set of players involved in most data initiatives, the ability to orchestrate resources and business objectives will be critical to success. A data initiative driven by the head of sales may quickly move beyond the siloed CRM or prospecting databases controlled by the sales department to include data feeds from other sources within and outside of the organization. Knowledge of underlying data technology, knowledge of sales metrics, knowledge of data mining and a host of other skills may be needed to get the most out of this type of initiative.



CompTIA

# BIG DATA INSIGHTS AND OPPORTUNITIES

## SECTION 4: IT CHANNEL PARTNER PERSPECTIVES OF BIG DATA

RESEARCH


SECOND ANNUAL • SEPTEMBER 2013

www.comptia.org

## Key Points

- CompTIA research indicates growing momentum with big data offerings among IT firms in the channel. Nearly 1 in 3 IT channel partners report providing big data application deployment or integration services. Over the next 12 months, an additional 14% expect to begin offering big data consulting or advisory services.

- IT firms plan to take a range of actions to better position themselves to capitalize on big data [and small data] opportunities. Top strategies include: investing in technical training, investigating partner programs and/or aligning with big data vendors, hiring staff with big data expertise and looking for partnership opportunities.

- As customers begin to test the big data waters, many early engagements will likely be project-based, such as application deployment, data center-cloud integration or custom development. Just as storage-as-a-service and BC/DR-as-a-service has made its way into customer engagements, there will be a time and place for big data-as-a-service solutions for certain customer segments.

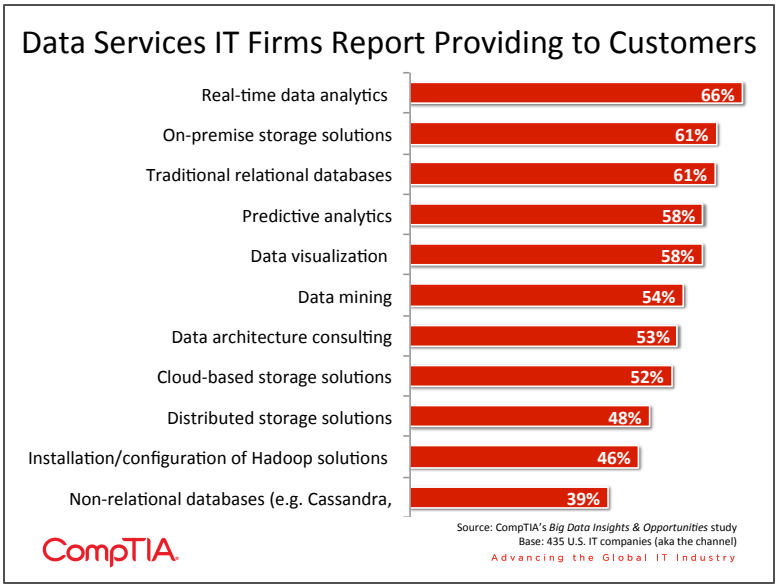
## Big Data and the Impact on the IT Channel

In the 2012 version of this study, IT companies with a channel presence were nearly equally divided between those who embraced the big data trend (54%) and those who believed it was premature to get too excited about big data business opportunities (aka "don't believe the hype" – 46%).

A year later, the same general sentiment holds. This should not be interpreted as a vote of no confidence in big data, but rather an acknowledgement of the uncertainty associated with trying to predict how a technology trend will play out. There have been plenty of "the next big thing" trends that never quite lived up to the hype. Lukewarm sentiment may also reflect concern with the selling and marketing of big data solutions – it may not be a matter of lack of opportunity, but rather the practical challenge of actually making money from the trend.

Nonetheless, the market has matured over the past 12 months. CompTIA research suggests a greater number of IT companies have taken action on the big data front in some way – be it providing new data-related services (such as those in the chart below), investing in training or evaluating new partners.

As noted in *Sections 1 and 2*, big data has both a specific definition, as well as a broader conceptual definition. Consequently, differing interpretations lead to differences in how customers and IT firms view their data-related activities. This should be taken into consideration when viewing incidence and adoption rates in the chart to the right.



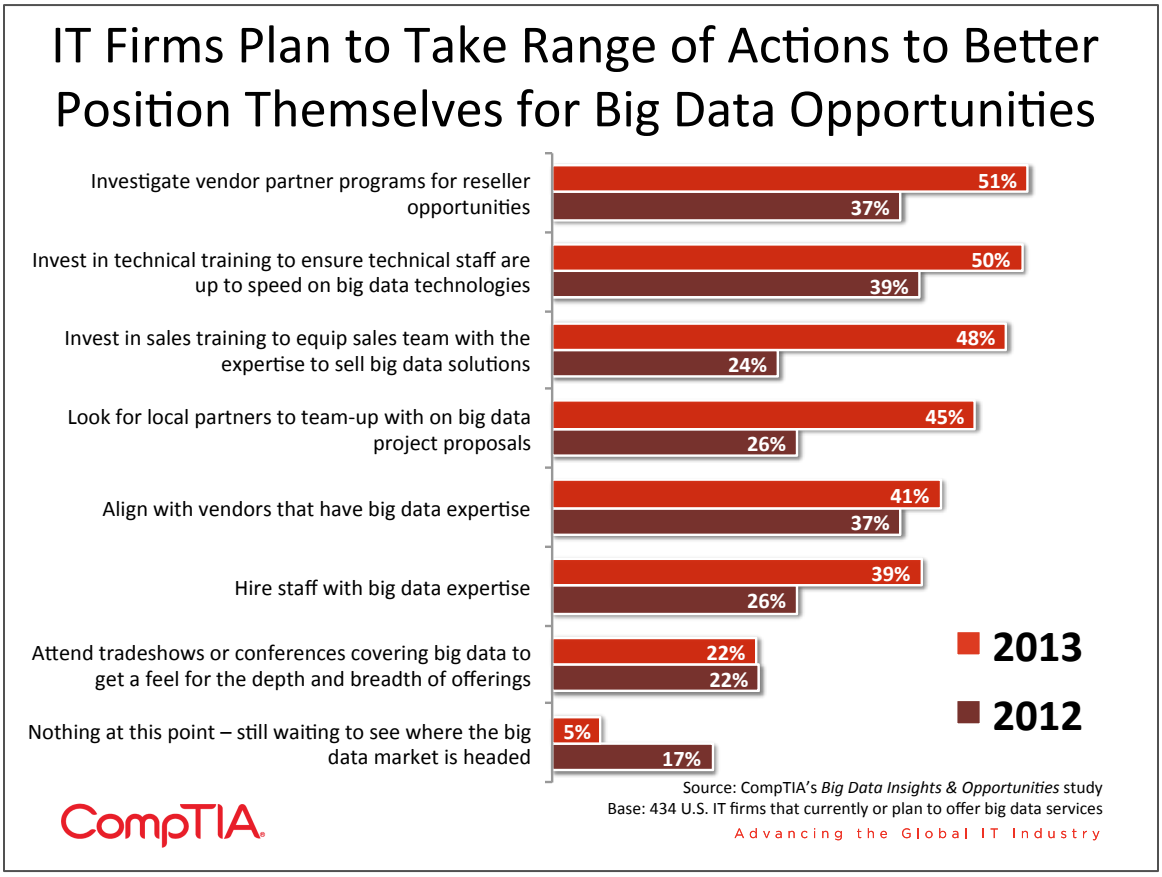

Additionally, because the big data market is still in its infancy, it is possible some IT firms that claim a big data offering haven't yet secured their first customer. Another factor, similar to the practice of "cloud washing," or attempting to re-brand a legacy product or service as a cloud offering, there will inevitably be some "big data washing." Some IT firms will be tempted to position certain products or services as a big data offering, even if it may be a bit of a stretch.

The research indicates there will be some additional up-take of big data solutions among IT firms, especially in the areas of non-relational databases – 28% of respondents signal plans to get involved in this part of the market over the next 12 months. A nearly identical percentage (27%) plan to add some type of Hadoop solution, such as deployment, configuration or management. Given the rapid pace of innovation and growth of distributed platforms, such as Hadoop, a move into the space makes sense. However, as noted in *Section 2*, Hadoop addresses a certain type of need under certain circumstances – it's still far from mainstream, so expectations should be kept in check.

Because adding a new line of business typically requires a substantial investment, it is not something taken lightly by channel partners, especially the smaller firms. Beyond expenditures in new technology or headcount, the opportunity costs of foregone investments must be considered. Moreover, because there are so many data-related business opportunities that support (e.g. storage virtualization) or are on the periphery of big data (e.g. business continuity and disaster recovery), many IT solution providers can indirectly benefit from the big data trend.

Among firms that do plan to pursue big data business opportunities, a range of strategies will be employed. Because each channel partner has a different set of capabilities, value propositions, customer bases and so on, it follows that there is no "one size, fits all" strategy for building a big data line of business. For more on building a big data practice, see CompTIA's channel training guide, *Easing Into Big Data*, http://www.comptia.org/training/business.aspx.

When viewing the research, a few common themes emerge for how IT firms plan to enter or expand in the big data market.



Three common elements of entering a new market include:

- What to sell?
- How to execute on the deliverable?
- How to sell it?

With 51% of channel firms indicating intent to investigate vendor partner programs, the data suggests many are well on their way to developing a strategy for "what to sell." Many established vendors (think IBM, HP, EMC, etc.) have rolled out new big data offerings or have begun to re-position existing offerings with a big data twist. Along side these major players sit a growing class of big data startups (think Cloudera, Hortonworks, GoodData, etc.). Whether established or emerging, vendors looking to grow market share for their products inevitably turn to channel partners. While ultimately mutually beneficial, channel relationships do come with costs (direct and/or indirect) to both vendors and their partners, so each will be looking to find the right fit.

Next, the ability to execute on the deliverable is a function of capabilities and expertise. If gaps exist, staff training or hiring new employees typically follows. The research indicates channel partners' immediate data-related training needs tilt towards technical training, which could range from the hardware side of big data, to the software side, to the services side, or all of the above. As noted throughout this report, big data has become somewhat of a proxy for many data-related topics, so the pursuit of technical training in this research shouldn't automatically be viewed as Hadoop training or something comparable. While some will need Hadoop training, others may benefit from lower-tier data-related training or training covering a topic on the periphery of big data. As new big data (and related) training and education content seems to materialize everyday, those looking to develop their skills have plenty of options. See *Section 3* of this report for examples of training and the accompanying certifications.

Lastly, sales, marketing and a go-to-market strategy helps to ensure the aforementioned investments actually generate revenue. Forty-eight percent of IT channel firms plan to pursue training in this area over the next 12 months, a doubling of the 2012 rate. This may be viewed as another validation of the market – IT firms are more bullish on customer demand for data services and therefore willing to invest in sales to capitalize on the opportunity. There could also be a timing effect at play. Last year, the research indicated firms were far more likely to seek technical data-related training. Perhaps now, these firms are feeling more confident about their data capabilities and are ready to pursue customers more aggressively.

As noted previously, selling data-related solutions will likely require new approaches. Technology vendors and IT solutions providers will need to be prepared for a multi-faceted customer engagement involving technical and business objective problem-solving with staff across many functional areas.

| Functional Area | Example of the Type of Need Driving Their Data Initiatives |
|---|---|
| CIO | Efficient and more effective data storage and management |
| CEO/President | Transparency and visibility throughout the organization |
| CFO | Predictive analytics and more robust forecasting |
| CMO | Fine-tuning audience segmentation for micro targeting of ad campaigns |
| Head of R&D | Faster, more effective experimentation |
| Heads of Sales | Supporting human decision-making with data-driven, algorithm-based decisions |

## Customer Big Data Needs: The Channel Perspective

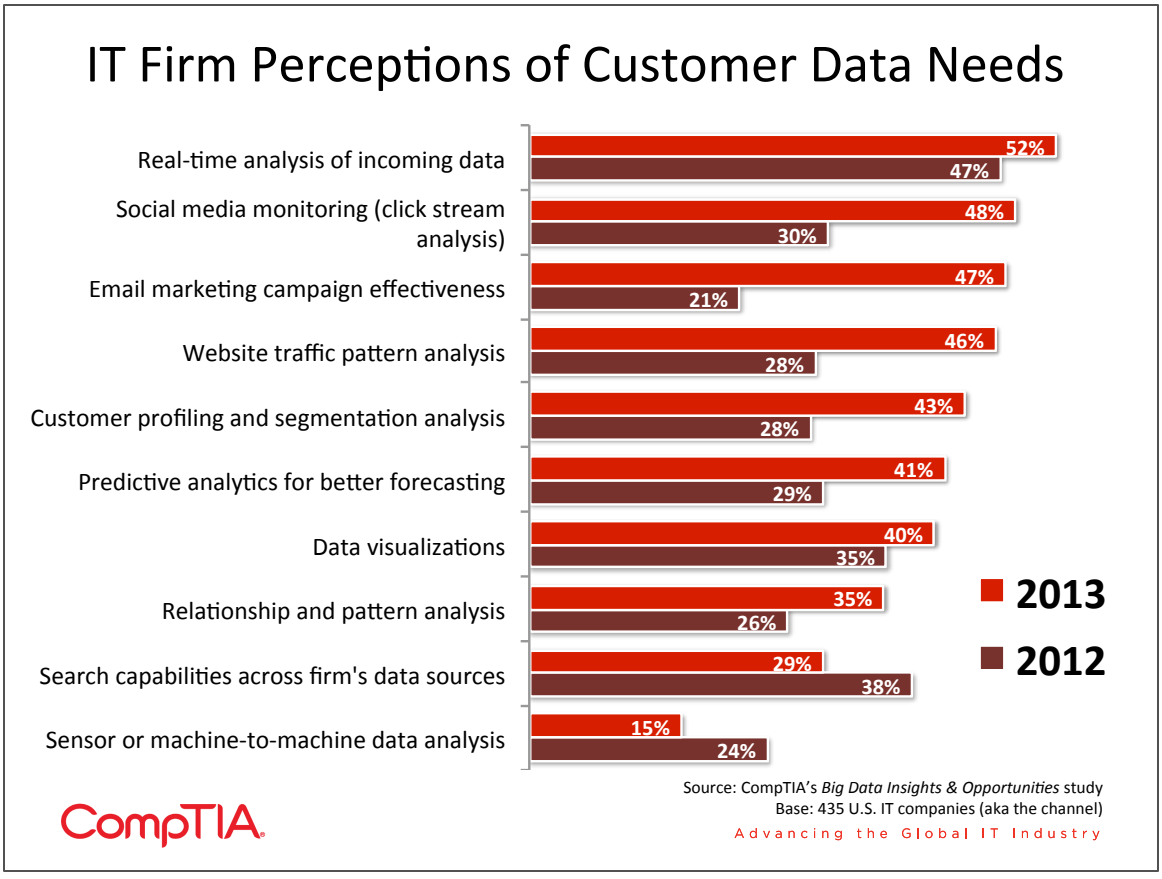
Savvy IT firms anticipate customer needs and proactively provide optimized solutions. When disconnects occur between perceived needs and actual customer needs, missed opportunities or inefficient customer engagements often follow.

According to the research, channel partners generally have a good read on customer data/big data needs. In some areas, such as the need for real-time analytics of incoming data, channel firms appear to be more in-sync with customers this year compared to last year. Improvement has also occurred in the area of anticipating the need for web analytics and customer segmentation and profiling. Overall, channel firms recognize more customer data needs this year compared to last year.

Despite the improvement, there may still be a couple of areas where channel partners could be underestimating customer demand.

- Search capabilities across a firm's data sources
- Predictive analytics
- End customer profiling and segmentation analysis
- Data visualization

Of course, this doesn't necessarily mean these needs should be the lead in a marketing campaign. Rather, it should serve as a reminder of the many nuances of the emerging big data market and the value of developing a true understanding of customer needs. As a reminder, big data and data initiatives in general should be thought of as a means to an end (e.g. faster decision making or higher sales conversion rates), rather than a technology investment that is the end in itself.



www.comptia.org