Big Data in Big Companies

Date: May 2013

Authored by:

Thomas H. Davenport
Jill Dyché

Introduction

Big data burst upon the scene in the first decade of the 21st century, and the first organizations to embrace it were online and startup firms. Arguably, firms like Google, eBay, LinkedIn, and Facebook were built around big data from the beginning. They didn't have to reconcile or integrate big data with more traditional sources of data and the analytics performed upon them, because they didn't have those traditional forms. They didn't have to merge big data technologies with their traditional IT infrastructures because those infrastructures didn't exist. Big data could stand alone, big data analytics could be the only focus of analytics, and big data technology architectures could be the only architecture.

Consider, however, the position of large, well-established businesses. Big data in those environments shouldn't be separate, but must be integrated with everything else that's going on in the company. Analytics on big data have to coexist with analytics on other types of data. Hadoop clusters have to do their work alongside IBM mainframes. Data scientists must somehow get along and work jointly with mere quantitative analysts.

In order to understand this coexistence, we interviewed 20 large organizations in the early months of 2013 about how big data fit in to their overall data and analytics environments. Overall, we found the expected co-existence; in not a single one of these large organizations was big data being managed separately from other types of data and analytics. The integration was in fact leading to a new management perspective on analytics, which we'll call "Analytics 3.0." In this paper we'll describe the overall context for how organizations think about big data, the organizational structure and skills required for it…etc. We'll conclude by describing the Analytics 3.0 era.

1: Big Data in Big Companies: How New?

Big data may be new for startups and for online firms, but many large firms view it as something they have been wrestling with for a while. Some managers appreciate the innovative nature of big data, but more find it "business as usual" or part of a continuing evolution toward more data. They have been adding new forms of data to their systems and models for many years, and don't see anything revolutionary about big data. Put another way, many were pursuing big data before big data was big.

When these managers in large firms are impressed by big data, it's not the "bigness" that impresses them. Instead it's one of three other aspects of big data: the lack of structure, the opportunities presented, and low cost of the technologies involved. This is consistent with the results from a survey of more than fifty large companies by NewVantage Partners in 2012. It found, according to the survey summary:

> *It's About Variety, not Volume: The survey indicates companies are focused on the variety of data, not its volume, both today and in three years. The most important goal and potential reward of Big Data initiatives is the ability to analyze diverse data sources and new data types, not managing very large data sets.*[i]

Firms that have long handled massive volumes of data are beginning to enthuse about the ability to handle a new type of data—voice or text or log files or images or video. A retail bank, for example, is getting a handle on its multi-channel customer interactions for the first time by analyzing log files. A hotel firm is analyzing customer lines with video analytics. A health insurer is able to better predict customer dissatisfaction by analyzing speech-to-text data from call center recordings. In short, these companies can have a much more complete picture of their customers and operations by combining unstructured and structured data.

There are also continuing—if less dramatic—advances from the usage of more structured data from sensors and operational data-gathering devices. Companies like GE, UPS, and Schneider National are increasingly putting sensors into things that move or spin, and capturing the resulting data to better optimize their businesses. Even small benefits provide a large payoff when adopted on a large scale. GE estimates that a 1% fuel reduction in the use of big data from aircraft engines would result in a $30 billion savings for the commercial airline industry over 15 years. Similarly, GE estimates that a 1% efficiency improvement in global gas-fired power plant turbines could yield a $66 billion savings in fuel consumption.[ii] UPS has achieved similarly dramatic savings (see the "Big Data at UPS" case study) through better vehicle routing.

2: Objectives for Big Data

Like many new information technologies, big data can bring about dramatic cost reductions, substantial improvements in the time required to perform a computing task, or new product and service offerings. Like traditional analytics, it can also support internal business decisions. The technologies and concepts behind big data allow organizations to achieve a variety of objectives, but most of the organizations we interviewed were focused on one or two. The chosen objectives have implications for not only the outcome and financial benefits from big data, but also the process—who leads the initiative, where it fits within the organization, and how to manage the project.

Cost Reduction from Big Data Technologies

Some organizations pursuing big data believe strongly that MIPS and terabyte storage for structured data are now most cheaply delivered through big data technologies like Hadoop clusters. One company's cost comparison, for example, estimated that the cost of storing one terabyte for a year was

$37,000 for a traditional relational database, $5,000 for a database appliance, and only $2,000 for a Hadoop cluster.[1] Of course, these figures are not directly comparable, in that the more traditional technologies may be somewhat more reliable and easily managed. Data security approaches, for example, are not yet fully developed in the Hadoop cluster environment.

[1] Paul Barth at NewVantage Partners supplied these cost figures.

Organizations that were focused on cost reduction made the decision to adopt big data tools primarily within the IT organization on largely technical and economic criteria. IT groups may want to involve some of your users and sponsors in debating the data management advantages and disadvantages of this kind of storage, but that is probably the limit of the discussion needed.

Big Data at UPS

UPS is no stranger to big data, having begun to capture and track a variety of package movements and transactions as early as the 1980s. The company now tracks data on 16.3 million packages per day for 8.8 million customers, with an average of 39.5 million tracking requests from customers per day. The company stores over 16 petabytes of data.

Much of its recently acquired big data, however, comes from telematics sensors in over 46,000 vehicles. The data on UPS package cars (trucks), for example, includes their speed, direction, braking, and drive train performance. The data is not only used to monitor daily performance, but to drive a major redesign of UPS drivers' route structures. This initiative, called ORION (On-Road Integrated Optimization and Navigation), is arguably the world's largest operations research project. It also relies heavily on online map data, and will eventually reconfigure a driver's pickups and drop-offs in real time. The project has already led to savings in 2011 of more than 8.4 million gallons of fuel by cutting 85 million miles off of daily routes. UPS estimates that saving only one daily mile driven per driver saves the company $30 million, so the overall dollar savings are substantial. The company is also attempting to use data and analytics to optimize the efficiency of its 2000 aircraft flights per day.

Cost reduction can also be a secondary objective after others have been achieved. Let's say, for example, that an organization's first goal was to innovate with new products and services from big data. After accomplishing that objective, it may want to examine how to do it less expensively. That was the case, for example, at GroupM, the media-buying subsidiary for the advertising conglomerate WPP.[iii] The company buys more media than any other organization in the world, and it uses big data tools to keep track of who's watching it on what screen. The only problem is that the GroupM has 120 offices around the world, and each office was taking its own approach—with its own technology—to big data analytics. If the organization allowed each office to implement its own big data tools, it would cost at least $1 million per site.

Instead of this highly decentralized approach, GroupM plans to offer centralized big data services out of its New York office. It will focus on 25 global markets, and expects that it will spend just over a third of the amount per site that the decentralized approach would have required. We expect to see many more of these consolidations in the next several years, as firms that allowed decentralized experimentation with big data attempt to rein in their costs.

We'll discuss other ways to drive value from big data later in Section 6.

Big Data at an International Financial Services Firm

For one multinational financial services institution, cost savings is not only a business goal, it's an executive mandate. The bank is historically known for its experimentation with new technologies, but after the financial crisis, it is focused on building its balance sheet and is a bit more conservative with new technologies. The current strategy is to execute well at lower cost, so the bank's big data plans need to fit into that strategy. The bank has several objectives for big data, but the primary one is to exploit "a vast increase in computing power on dollar-for-dollar basis." The bank bought a Hadoop cluster, with 50 server nodes and 800 processor cores, capable of handling a petabyte of data. IT managers estimate an order of magnitude in savings over a traditional data warehouse. The bank's data scientists—though most were hired before that title became popular—are busy taking existing analytical procedures and converting them into the Hive scripting language to run on the Hadoop cluster.

According to the executive in charge of the big data project, "This was the right thing to focus on given our current situation. Unstructured data in financial services is somewhat sparse anyway, so we are focused on doing a better job with structured data. In the near to medium term, most of our effort is focused on practical matters—those where it's easy to determine ROI—driven by the state of technology and expense pressures in our business. We need to self-fund our big data projects in the near term. There is a constant drumbeat of 'We are not doing "build it and they will come'—we are working with existing businesses, building models faster, and doing it less expensively. This approach is more sustainable for us in the long run. We expect we will generate value over time and will have more freedom to explore other uses of big data down the road."

Time Reduction from Big Data

The second common objective of big data technologies and solutions is time reduction. Macy's merchandise pricing optimization application provides a classic example of reducing the cycle time for complex and large-scale analytical calculations from hours or even days to minutes or seconds. The department store chain has been able to reduce the time to optimize pricing of its 73 million items for sale from over 27 hours to just over 1 hour. Described by some as "big data analytics," this capability set obviously makes it possible for Macy's to re-price items much more frequently to adapt to changing conditions in the retail marketplace. This big data analytics application takes data out of a Hadoop cluster and puts it into other parallel computing and in-memory software architectures. Macy's also says it achieved 70% hardware cost reductions. Kerem Tomak, VP of Analytics at Macys.com, is using similar approaches to time reduction for marketing offers to Macy's customers (see the, "Big Data at Macys.com," case study). He notes that the company can run a lot more models with this time savings:

> *Generating hundreds of thousands of models on granular data versus only 10, 20 or the 100 that we used to be able to run on aggregate data is really the key difference between what we can do now and what we will be able to do with high performance computing.*[iv]

Tomak also makes extensive use of visual analytics tools for his big data results, which is common with big data.

Another key objective involving time reduction is to be able to interact with the customer in real time, using analytics and data derived from the customer experience. If the customer has "left the building," targeted offers and services are likely to be much less effective. This means rapid data capture, aggregation, processing, and analytics (see the "Big Data at Caesars Entertainment" case study).

Developing New Big Data-Based Offerings

One of the most ambitious things an organization can do with big data is to employ it in developing new product and service offerings based on data. Many of the companies that employ this approach are online firms, which have an obvious need to employ data-based products and services. The best example may be LinkedIn, which has used big data and data scientists to develop a broad array of product offerings and features, including People You May Know, Groups You May Like, Jobs You May Be Interested In, Who's Viewed My Profile, and several others. These offerings have brought millions of new customers to LinkedIn.

Another strong contender for the best at developing products and services based on big data is Google. This company, of course, uses big data to refine its core search and ad-serving algorithms. Google is constantly developing new products and services that have big data algorithms for search or ad

Big Data at Caesars Entertainment

Caesars (formerly Harrah's) Entertainment has long been a leader in the use of analytics, particularly in the area of customer loyalty, marketing, and service. Today, Caesars is augmenting these traditional analytics capabilities with some big data technologies and skills. The primary objective of exploring and implementing big data tools is to respond in real time for customer marketing and service.

For example, the company has data about its customers from its Total Rewards loyalty program, web clickstreams, and from real-time play in slot machines. It has traditionally used all those data sources to understand customers, but it has been difficult to integrate and act on them in real time, while the customer is still playing at a slot machine or in the resort.

In order to pursue this objective, Caesars has acquired both Hadoop clusters and open-source and commercial analytics software. It has also added some data scientists to its analytics group.

There are other goals for the big data capabilities as well. Caesars pays fanatical attention—typically through human observation—to ensuring that its most loyal customers don't wait in lines. With video analytics on big data tools, it may be able to employ more automated means for spotting service issues involving less frequent customers. Caesars is also beginning to analyze mobile data, and is experimenting with targeted real-time offers to mobile devices.

placement at the core, including Gmail, Google Plus, Google Apps, etc. Google even describes the self-driving car as a big data application.[v] Some of these product developments pay off, and some are discontinued, but there is no more prolific creator of such offerings than Google.

There are many other examples of this phenomenon, including among firms outside of the online industry. Among the companies we interviewed for this research, GE is the most prominent creator of new service offerings based on big data. GE is primarily focused on optimizing the service contracts and maintenance intervals for industrial products (see the "Big Data at GE" case study).

But there are many other examples in businesses with a substantial data component. Verizon Wireless, Sprint, and T-Mobile all have already, or are in the process of, selling services based on usage and location data from mobile devices. Verizon's Precision Market Insights offerings, for example, evaluate the effectiveness of outdoor media locations, retail store locations, and venue audiences. Netflix created the well-known Netflix prize for the data science team that could optimize the company's movie recommendations for customers, and is now using big data to help in the creation of proprietary content, including its successful "House of Cards" series.[vi] The testing firm Kaplan uses its big data to begin advising customers on effective learning and test-preparation strategies. These companies' big data efforts are directly focused on products, services, and customers.

This has important implications, of course, for the organizational locus of big data and the processes and pace of new product development. If an organization is serious about product and service generation with big data, it will need to create a broad program or initiative for doing so—a set of tools, technologies, and people who are good at big data manipulation. GE's new center for software and data analytics qualifies as such a program.

Supporting Internal Business Decisions

The primary purpose behind traditional, "small data" analytics was to support internal business decisions. What offers should be presented to a customer? Which customers are most likely to stop being customers soon? How much inventory should be held in the warehouse? How should we price our products?

These types of decisions employ big data when there are new, less structured data sources that can be applied to the decision. For example, any data that can shed light on customer satisfaction is helpful, and much data from customer interactions is unstructured (see the "Big Data at United Healthcare" case study.)

Three major banks we interviewed - Wells Fargo, Bank of America, and Discover - are also using big data to understand aspects of the customer relationship that they couldn't previously get at. In that industry—as well as several others, including retail—the big challenge is to understand multi-channel customer relationships. They are monitoring customer "journeys" through the tangle of websites, call centers, tellers, and other branch personnel to understand the paths that customers follow through the bank, and how those paths affect attrition or the purchase of particular financial services.

The data sources on multi-channel customer journeys are unstructured or semi-structured. They include website clicks, transaction records, bankers' notes, and voice recordings from call centers. The volumes are quite large—12 billion rows of data for one of the banks. All three banks are beginning to better

Big Data at United Healthcare

United Healthcare, like many large organizations pursuing big data, has been focused on structured data analysis for many years, and even advertises its analytical capabilities to consumers ("Health in Numbers"). Now, however, it is focusing its analytical attention on unstructured data—in particular, the data on customer attitudes that is sitting in recorded voice files from customer calls to call centers. The level of customer satisfaction is increasingly important to health insurers, because consumers increasingly have choice about what health plans they belong to. Service levels are also being monitored by state and federal government groups, and published by organizations such as *Consumer Reports*.

In the past, that valuable data from calls couldn't be analyzed. Now, however, United is turning the voice data into text, and then analyzing it with "natural language processing" software. The analysis process can identify—though it's not easy, given the vagaries of the English language—customers who use terms suggesting strong dissatisfaction. A United representative can then make some sort of intervention—perhaps a call exploring the nature of the problem. The decision being made is the same as in the past—how to identify a dissatisfied customer—but the tools are different.

To analyze the text data, United Healthcare uses a variety of tools. The data initially goes into a "data lake" using Hadoop and NoSQL storage, so the data doesn't have to be normalized. The natural language processing—primarily a "singular value decomposition", or modified word count—takes place on a database appliance. A variety of other technologies are being surveyed and tested to assess their fit within the "future state architecture. United also makes use of interfaces between its statistical analysis tools and Hadoop.

The work to put the customer satisfaction data, along with many other sources of customer data, into a customer data warehouse and analyze it is being led by Mark Pitts, who is based in the Finance organization. However, several other functions and units of United, including its Optum business specializing in selling data and related services to healthcare organizations, are participating. Pitt's team includes both conventional quantitative analysts and data scientists with strong IT and data management skills.

understand common journeys, describing them with segment names, ensuring that the customer interactions are high-quality, identifying reasons for attrition, and correlating journeys with customer opportunities and problems. It's a complex set of problems and decisions to analyze, but the potential payoff is high.

Business decisions with big data can also involve other traditional areas for analytics such as supply chains, risk management, or pricing. The factor that makes these big data problems, rather than small, is the use of external data to improve the analysis. In supply chain decisions, for example, companies are increasingly using external data to measure and monitor supply chain risks. External sources of supplier

data can furnish information on suppliers' technical capabilities, financial health, quality management, delivery reliability, weather and political risk, market reputation, and commercial practices. The most advanced firms are monitoring not only their own suppliers, but their suppliers' suppliers.

3: Big Data's Moving Parts

No single business trend in the last decade has as much potential impact on incumbent IT investments as big data. Indeed big data promises—or threatens, depending on how you view it—to upend legacy technologies at many big companies. As IT modernization initiatives gain traction and the accompanying cost savings hit the bottom line, executives in both line of business and IT organizations are getting serious about the technology solutions that are tied to big data.

Companies are not only replacing legacy technologies in favor of open source solutions like Apache Hadoop, they are also replacing proprietary hardware with commodity hardware, custom-written applications with packaged solutions, and decades-old business intelligence tools with data visualization. This new combination of big data platforms, projects, and tools is driving new business innovations, from faster product time-to-market to an authoritative—finally!—single view of the customer to custom-packaged product bundles and beyond.

The Big Data Stack

As with all strategic technology trends, big data introduces highly specialized features that set it apart from legacy systems. Figure 1 illustrates the typical components of the big data stack.

Each component of the stack is optimized around the large, unstructured and semi-structured nature of big data. Working together, these moving parts comprise a holistic solution that's fine-tuned for specialized, high-performance processing and storage.

Storage:

Storing large and diverse amounts of data on disk is becoming more cost-effective as the disk technologies become more commoditized and efficient. Companies like EMC sell storage solutions that allow disks to be added quickly and cheaply, thereby scaling storage in lock step with growing data volumes. Indeed, many big company executives see Hadoop as a low-cost alternative for the archival and quick retrieval of large amounts of historical data.

Platform Infrastructure:

The big data "platform" is typically the collection of functions that comprise high-performance processing of big data. The platform includes capabilities to integrate, manage, and apply sophisticated computational processing to the data. Typically, big data platforms include a Hadoop (or similar open-source project) foundation. Hadoop was designed and built to optimize complex manipulation of large amounts of data while vastly exceeding the price/performance of traditional databases. Hadoop is a unified storage and processing environment that is highly scalable to large and complex data volumes.

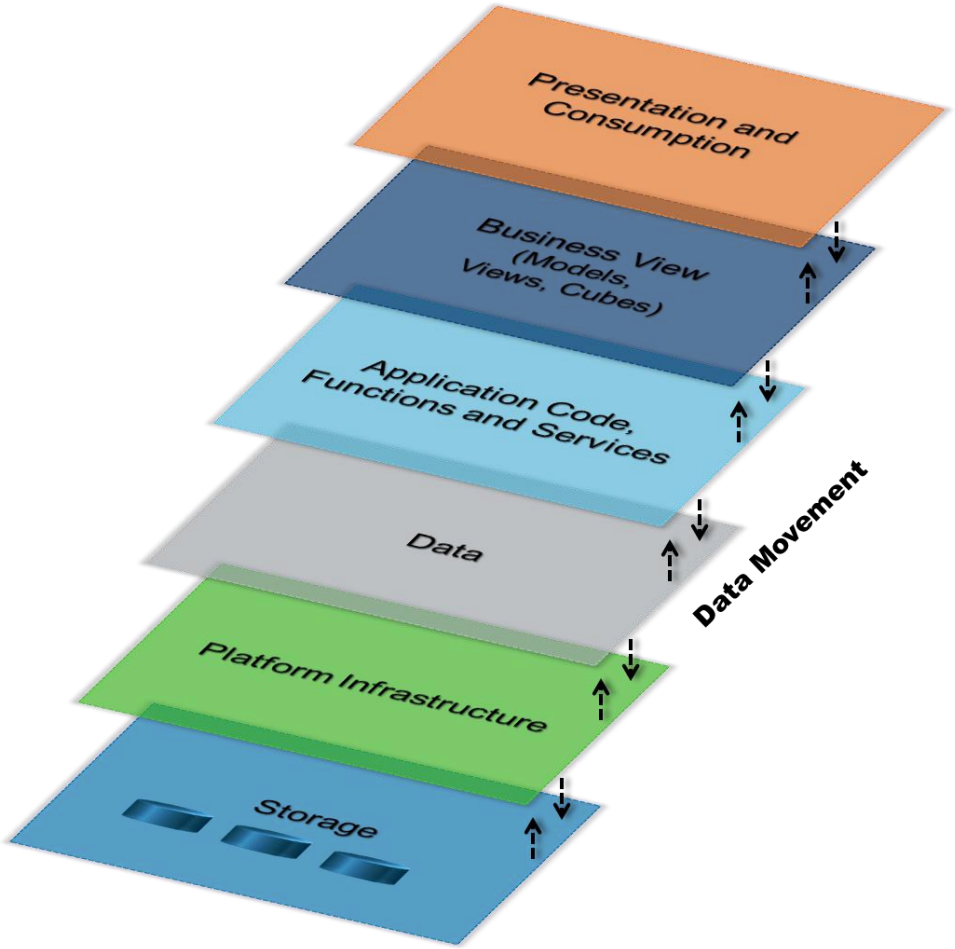

Figure 1: The Big Data Stack

You can think of it as big data's execution engine. The Lead Information Architect in a large property and casualty insurer (see the "Big Data at a Top 5 Property and Casualty Insurer" case study) knows this firsthand:

> *"We knew there were a lot of opportunities for Hadoop when we started," says the Lead Information Architect. "So we loaded some data into Hadoop. After making some quick calculations, we realized that the data we'd just loaded would have exceeded the capacity of our data warehouse. It was impressive."*

In the new world of big data, open source projects[vii] like Hadoop have become the de facto processing platform for big data. Indeed, the rise of big data technologies has meant that the conversation around analytics solutions has fundamentally changed. Companies unencumbered with legacy data warehouses (many recent high-tech startups among them) can now leverage a single Hadoop platform to segregate complex workloads and support a variety of usage scenarios from complex mathematical computations to ad hoc visualizations.

Data:

The expanse of big data is as broad and complex as the applications for it. Big data can mean human genome sequences, oil well sensors, cancer cell behaviors, locations of products on pallets, social media interactions, or patient vital signs, to name a few examples. The data layer in the stack implies that data is a separate asset, warranting discrete management and governance.

To that end, a 2013 survey of data management professionals[viii] found that of the 339 companies responding, 71 percent admitted that they "have yet to begin planning" their big data strategies. The respondents cited concerns about data quality, reconciliation, timeliness, and security as significant barriers to big data adoption.

> *"It is challenging, but critically important, to prioritize the type of analytics we do while simultaneously integrating data, technologies and other resources" shared Allen Naidoo, Vice President for Advanced Analytics at Carolinas Health Care. Indeed, the health care provider has plans on adding genetics data to its big data roadmap as soon as it formalizes some of the more complex governance and policy issues around the data.*

Application Code, Functions, and Services:

Just as big data varies with the business application, the code used to manipulate and process the data can vary. Hadoop uses a processing engine called MapReduce to not only distribute data across the disks, but to apply complex computational instructions to that data. In keeping with the high-performance capabilities of the platform, MapReduce instructions are processed in parallel across various nodes on the big data platform, and then quickly assembled to provide a new data structure or answer set.

An example of a big data application in Hadoop might be to "calculate all the customers who like us on social media." A text mining application might crunch through social media transactions, searching for words such as "fan," "love," "bought," or "awesome" and consolidate a list of key influencer customers.

Business View:

Depending on the big data application, additional processing via MapReduce or custom Java code might be used to construct an intermediate data structure, such as a statistical model, a flat file, a relational table, or a cube. The resulting structure may be intended for additional analysis, or to be queried by a traditional SQL-based query tool. This business view ensures that big data is more consumable by the tools and the knowledge workers that already exist in an organization.

One Hadoop project called "Hive" enables raw data to be re-structured into relational tables that can be accessed via SQL and incumbent SQL-based toolsets, capitalizing on the skills that a company may already have in-house.

Presentation and Consumption:

One of the more profound developments in the world of big data is the adoption of so-called data visualization. Unlike the specialized business intelligence technologies and unwieldy spreadsheets of

yesterday, data visualization tools allow the average business person to view information in an intuitive, graphical way.

For instance, if a wireless carrier wants better information on its network, looking in particular for patterns in dropped calls, it could assemble a complicated spreadsheet full of different columns and figures. Alternatively, it could deploy an easy-to-consume, graphical trend report to its field service staff, like the one shown in Figure 2.

This data visualization displays three different views of the data. The first shows call drops by region grouped by network generation. The second shows how each hour the distribution of dropped calls is different. The third shows a higher percentage of dropped calls in the 4G network around the call start hour of 17:00. This kind of information might prompt a network operator to drill down further and discover the root causes of problems on the network, and which high-value customers might be affected by them.

Cell Tech (last)	2G		3G			4G		
Subscriber Region	Call Duration (minutes)	% Call drops	Calls duration	Call Duration (minutes)	% Call drops	Calls duration	Call Duration (minutes)	% Call drops
Midwest	3,566.9	20.7%	10,700	17,353.9	23.3%	1,010	1,587.4	28.2%
Mtn West	1,050.8	16.2%	3,470	5,787.7	26.8%	370	545.7	36.5%
New England	2,355.0	25.7%	9,310	14,633.3	20.4%	920	1,649.6	21.4%
Pacific	7,321.6	25.0%	27,834	41,972.6	23.8%	2,891	4,454.5	26.4%
South	10,454.2	27.6%	35,999	55,717.4	25.3%	3,797	5,962.6	26.1%
Southwest	1,274.0	16.1%	4,622	7,000.3	26.7%	498	604.5	31.4%

Figure 2: Data Visualization at a Wireless Carrier

Such a visualization can be pulled by the network operator onto her desktop PC, or pushed to the mobile device of a service technician in the field, thereby decreasing time-to-resolution for high-impact trouble tickets. And it can be done in a fraction of the time it used to take finding, accessing, loading, and consolidating the data from myriad billing and customer systems.

Data visualizations, although normally highly appealing to managerial users, are more difficult to create when the primary output is a multivariate predictive model; humans have difficulty understanding visualizations in more than two dimensions. Some data visualization tools now select the most appropriate visual display for the type of data and number of variables. Of course, if the primary output of a big data analysis is an automated decision, there is no need for visualization.

Changing the Vocabulary

The brave new world of big data not only upends the traditional technology stack when it comes to high-performance analytics, it challenges traditional ways of using and accessing data. At the very least, it starts to change the company's vocabulary about information and its role in analytics.

> *"Big data wasn't really a commonly-known term when I first started," explains Karem Tomak of Macys.com. "I knew we needed a cheap, scalable platform. We did a proof-of-concept comparing a Hadoop cluster against a DB2 cluster. Hadoop was much faster." With all that data—and a 50 percent year-over-year growth rate—it's more than likely that the business demand for big data at Macys.com will only increase.*

As we discuss in the next section, that can have a dramatic effect on development organizations and delivery skills. And as we'll see in Section 5 of this report, it can also have an impact on incumbent analytics technologies.

Big Data at a Top 5 Property and Casualty Insurer

Started in 1922 by a handful of military officers who offered to insure each other's vehicles when no one else would, the insurer has become a financial services powerhouse, offering a broad range of insurance and banking services to military members and their families. The size of its customer base and breadth of products make big data a natural next step in the company's already-advanced technology portfolio.

Consistently named one of the country's best places to work and lauded for its many customer service awards, the insurer has made understanding customer behaviors and preferences core to its mission. "We have a strategy of continuing our evolution as a 'relationship' company," explained the Lead Information Architect in the insurer's BI Lab and one of the visionaries behind the company's big data roadmap. "This means taking into account as many data sources as possible, and being able to harness as many new types of data as we need."

In addition to cultivating a deeper view into customers' product needs and service preferences, the insurer is using a new crop of big data solutions for fraud detection—monitoring data patterns to pinpoint "points of compromise"—using telematics data to provide in-vehicle service, and sensory telemetry information for its mobile apps.

4. Organizational Structures and Skills

As with technology architectures, organizational structures and skills for big data in big companies are evolving and integrating with existing structures, rather than being established anew. No organization we interviewed has established an entirely separate organization for big data; instead, existing analytics or technology groups have added big data functions to their missions. Some do not mark any change in organizational structure or skills at all with big data, stating that they have been analyzing large quantities of data for many years. Others simply say that they are adding data science capabilities to their existing portfolios.

Organizational Structures for Big Data

The most likely organizational structures to initiate or accommodate big data technologies are either existing analytics groups (including groups with an "operations research" title), or innovation or architectural groups within IT organizations. In many cases these central services organizations are aligned in big data initiatives with analytically-oriented functions or business units—marketing, for example, or the online businesses for banks or retailers (see the "Big Data at Macys.com" case study). Some of these business units have IT or analytics groups of their own. The organizations whose approaches seemed most effective and likely to succeed had close relationships between the business groups addressing big data and the IT organizations supporting them.

Big Data Skill Scarcity

In terms of skills, most of these large firms are augmenting—or trying to augment—their existing analytical staffs with data scientists who possess a higher level of IT capabilities, and the ability to manipulate big data technologies specifically—compared to traditional quantitative analysts. These might include natural language processing or text mining skills, video or image analytics, and visual analytics. Many of the data scientists are also able to code in scripting languages like Python, Pig, and Hive. In terms of backgrounds, some have Ph.D.s in scientific fields; others are simply strong programmers with some analytical skills. Many of our interviewees questioned whether a data scientist could possess all the needed skills, and were taking a team-based approach to assembling them.

A key skill involves explaining big data outcomes to executives—in visual displays or verbal narrative. Even among data scientists, several interviewees commented that their quantitative people need to "tell a story with data," and relate well to decision-makers. Several representatives in firms we interviewed pointed out that their analytics people are also required to devote considerable time to change management issues. As prescriptive analytical models are embedded into key operational processes, someone needs to work with front-line workers and process owners to bring about needed changes in roles, process designs, and skills.

Big Data at Macys.com

Macys.com is considered the equivalent of a single store at the giant retailer's structure, but it's growing at a 50% annual rate—faster than any other part of the business. The division's management is very oriented to and knowledgeable about IT, data, and analytical decisions. Like other online retailers, Macys.com is heavily focused on customer-oriented analytical applications involving personalization, ad and email targeting, and search engine optimization. Within the Macys.com analytics organization, the "Customer Insights" group addresses these issues, but it also has a "Business Insights" group (focused primarily on supporting and measuring activity around the marketing calendar) and a "Data Science" organization. The latter addresses more leading-edge quantitative techniques involving data mining, marketing, and experimental design.

Macys.com utilizes a variety of leading-edge technologies for big data, most of which are not used elsewhere within the company. They include open-source tools like Hadoop, R, and Impala, as well as purchased software such as SAS, IBM DB2, Vertica, and Tableau. Analytical initiatives are increasingly a blend of traditional data management and analytics technologies, and emerging big data tools. The analytics group employs a combination of machine learning approaches and traditional hypothesis-based statistics.

Kerem Tomak, who heads the analytics organization at Macys.com, argues that it's important not to pursue big data technology for its own sake. "We are very ROI-driven, and we only invest in a technology if it solves a business problem for us," he noted. Over time there will be increasing integration between Macys.com and the rest of Macy's systems and data on customers, since Tomak and his colleagues believe that an omnichannel approach to customer relationships is the right direction for the future.

It goes almost without saying that the skills, processes, and tools necessary to manage exploding amounts of non-standard data will become ever more scarce and important. For the most part, the companies we interviewed feel substantially less urgency than the startups we have encountered with regard to data science talent. For some, however, the talent shortage is beginning to bite.

The most active recruiter of data scientists among the companies we interviewed is GE, which has an objective of recruiting roughly 400 of them, and has already hired or transferred in from elsewhere in GE about half that many. Although GE has had considerable success in recruiting data scientists, it is also creating an internally-developed training program for them. It also occasionally has challenges recruiting data scientists who are familiar with the specific data issues around its industrial products, e.g., turbine sensor data.

Big Data at Bank of America

Given Bank of America's large size in assets (over $2.2 trillion in 2012) and customer base (50 million consumers and small businesses), it was arguably in the big data business many years ago. Today the bank is focusing on big data, but with an emphasis on an integrated approach to customers and an integrated organizational structure. It thinks of big data in three different "buckets"—big transactional data, data about customers, and unstructured data. The primary emphasis is on the first two categories.

With a very large amount of customer data across multiple channels and relationships, the bank historically was unable to analyze all of its customers at once, and relied on systematic samples. With big data technology, it can increasingly process and analyze data from its full customer set.

Other than some experiments with analysis of unstructured data, the primary focus of the bank's big data efforts is on understanding the customer across all channels and interactions, and presenting consistent, appealing offers to well-defined customer segments. For example, the Bank utilizes transaction and propensity models to determine which of its primary relationship customers may have a credit card, or a mortgage loan that could benefit from refinancing at a competitor. When the customer comes online, calls a call center, or visits a branch, that information is available to the online app, or the sales associate to present the offer. The various sales channels can also communicate with each other, so a customer who starts an application online but doesn't complete it, could get a follow-up offer in the mail, or an email to set up an appointment at a physical branch location.

A new program of "BankAmeriDeals," which provides cash-back offers to holders of the bank's credit and debit cards based on analyses of where they have made payments in the past. There is also an effort to understand the nature of and satisfaction from customer journeys across a variety of distribution channels, including online, call center, and retail branch interactions.

The bank has historically employed a number of quantitative analysts, but for the big data era they have been consolidated and restructured, with matrixed reporting lines to both the a central analytics group and to business functions and units. The consumer banking analytics group, for example, made up of the quantitative analysts and data scientists, reports to Aditya Bhasin, who also heads Consumer Marketing and Digital Banking. It is working more closely with business line executives than ever before.

Several companies also mentioned the need for combining data scientist skills with traditional data management virtues. Solid knowledge of data architectures, metadata, data quality and correction processes, data stewardship and administration dashboards, master data management hubs, matching algorithms, and a host of other data-specific topics are important for firms pursuing big data as a long-term strategic differentiator.

> *"We're building data governance into our big data processes," says* the Lead Information Architect at a top 5 property and casualty insurer. *"We're adding metadata and assigning data classification levels to understand usage and consumption. We're dealing with new types of data we haven't dealt with before. And we're calling out the features and meta-content that needs to be captured along the way."*

Not surprisingly, however, respondents on the "business side" (as opposed to IT) did not mention these data management capabilities, and in fact sometimes complained when IT groups tried to impose them.

Of course, the need for business-driven policy making and oversight of information transcends even the most sophisticated and well-planned big data programs. The most governance-conscious firms will develop governance approaches that cut across all data types. "Our data governance efforts will pervade our environment," confirms the Executive Director of BI Delivery and Governance at a top 5 property and casualty insurer. "It applies to our data warehouses, our marts, and even our operational systems. After all, we consider data a corporate asset—and we treat it that way, no matter what the application."

Indeed, many of the challenges for big data center on the data itself. It turns out that these challenges are in fact well-known ones, now newly veiled by the big data conversation, though no less important than before.

Data-Savvy Leadership

In 2011 a widely-read McKinsey report on big data[ix] cited a company's "data-driven mind-set" to be a key indicator of big data's value to companies. The report gauged corporate cultures of fact-based decision making (as opposed to gut feel) as an important indicator of big data's value potential. The report also argued that in the U.S. alone more than 1.5 million data-savvy managers would be needed to lead their organizations' big data initiatives.

But effective managers in leading firms have already deduced how big data will drive value for their companies, using stories of big data successes to justify their own efforts. Likewise they use stories of missteps and false starts to establish solid business cases and firm up their planning and deployment strategies. As the McKinsey study noted, "Many pioneering companies are already using big data to create value, and others need to explore how they can do the same if they are to compete."

The executives we interviewed for this report are part of the latter group who have convinced their leadership and constituents that launching big data is not only worthwhile, but worth more than it costs. Many have responsibilities not only for big data and analytics, but for other functions as well.

> *"We're making a big bet on big data," says Bill Ruh from GE. "With that said, the pilot projects we've put out there have solved some big problems already. Our early proof-points were important. Now we're moving forward with even more complex problem sets. We're making this a part of everything we do."*

5. Integrating Analytics Environments

In their constant quest to understand a patient's journey across the continuum of care, health care providers are eyeing big data technologies to drive the patient lifecycle, from an initial physician encounter and diagnosis through rehabilitation and follow-up. Such lifecycle management capabilities include patient transactions—social media interactions, radiology images, and pharmacy prescriptions among them—that can populate and enrich a patient's health record. This data can then be stored in HDFS, repopulated into the operational systems, or prepared for subsequent analytics via a data warehouse or mart.

Figure 3 illustrates a simple big data technology environment.

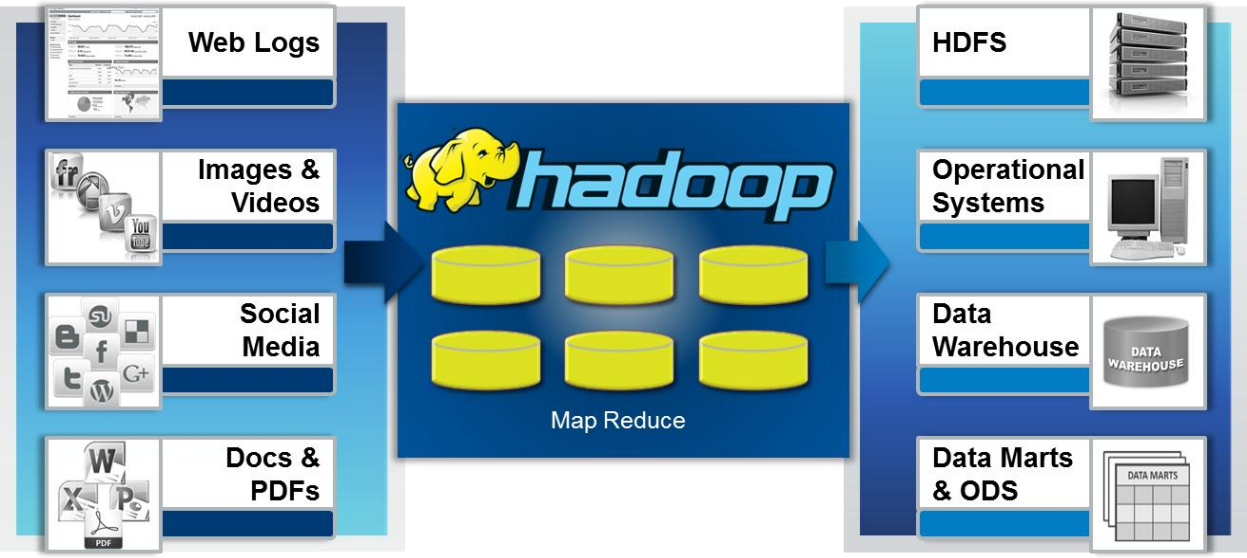

Figure 3: A Big Data Technology Ecosystem

Note that in the example the data sources themselves are heterogeneous, involving more diverse unstructured and semi-structured data sets like emails, web logs, or images. These data sources are increasingly likely to be found outside of the company's firewall. The big companies adopting production-class big data environments need faster and lower-cost ways to process large amounts of atypical data. Think of the computing horsepower needed by energy companies to process data streaming from smart meters, or by retailers tracking in-store smartphone navigation paths, or LinkedIn's reconciliation of millions of colleague recommendations.

Or consider a gaming company's ability to connect consumers with their friends via on-line video games.

"Before big data our legacy architecture was fairly typical," an on-line gaming executive explained to us. "Like most companies we had data warehouses and lots of ETL programs, and our data was very latent. And that meant that our analytics were very reactive."

The gaming company revamped not only its analytics technology stack, but the guiding principles on which it processed its data, stressing business relevance and scalability. The IT group adopted Hadoop and began using machine learning and advanced analytical algorithms to drive better prediction, thus optimizing customer offers and pricing.

"Once we were able to really exploit big data technology we could then focus on the gamer's overall persona," the executive said. "This allowed all the data around the gamer to be more accurate, giving us a single identity connecting the gamer to the games, her friends, the games her friends are playing, her payment and purchase history, and her play preferences. The data is the glue that connects everything."

Hadoop offers these companies a way to not only ingest the data quickly, but to process and store it for re-use. Because of its superior price-performance, some companies are even betting on Hadoop as a data warehouse replacement, acquiring SQL extensions in order to make big data more consumable for business users. Then again, many big companies have already invested millions in incumbent analytics environments, and they have no plans on replacing them anytime soon.

What Most Companies Do Today

The classic analytics environment at most big companies includes the operational systems that serve as the sources for data; a data warehouse or collection of federated data marts which house and—ideally—integrate the data for a range of analysis functions; and a set of business intelligence and analytics tools that enable decisions from the use of ad hoc queries, dashboards, and data mining. Figure 4 illustrates the typical big company data warehouse ecosystem.

Indeed, big companies have invested tens of millions of dollars in hardware platforms, databases, ETL (extraction, transformation and loading) software, BI dashboards, advanced analytics tools, maintenance contracts, upgrades, middleware, and storage systems that comprise robust, enterprise-class data warehouse environments.

In the best cases, these environments have helped companies understand their customer purchase and behavior patterns across channels and relationships, streamlining sales processes, optimizing product pricing and packaging, and driving more relevant conversations with prospects, thereby enhancing their brands. In the worst cases, companies have over-invested in these technologies, with many unable to recoup their investments in analytics, and viewing their data warehouse infrastructures as sunk costs with marginal business value.

The more mature a company's analytics environment, the more likely it is that it represents a combination of historical successes and failures. Best practice organizations approach BI and analytics not as a single project focused on a centralized platform, but as a series of business capabilities

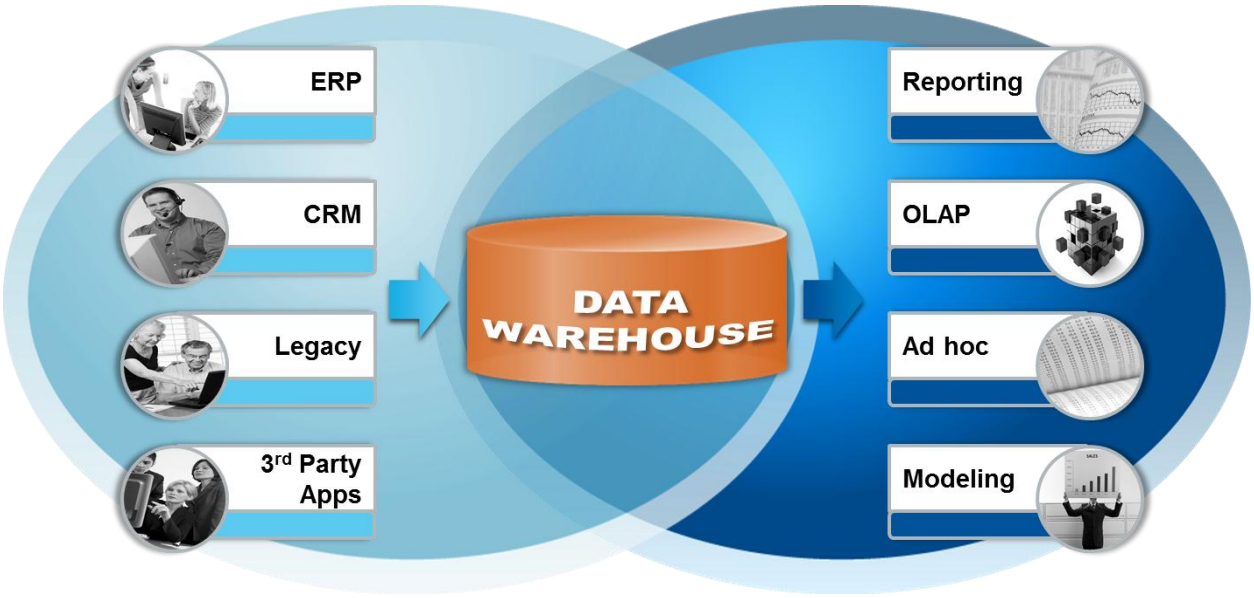

Figure 4: A Typical Data Warehouse Environment

deployed over time, exploiting a common infrastructure and re-usable data. As we discussed in Section 1, big data introduces fresh opportunities to expand this vision, and deploy new capabilities that incumbent systems aren't optimized to handle.

Putting the Pieces Together

Big companies with large investments in their data warehouses have neither the resources nor the will to simply replace an environment that works well doing what it was designed to do. At the majority of big companies, a coexistence strategy that combines the best of legacy data warehouse and analytics environments with the new power of big data solutions is the best of both worlds

Many companies continue to rely on incumbent data warehouses for standard BI and analytics reporting, including regional sales reports, customer dashboards, or credit risk history. In this new environment, the data warehouse can continue with its standard workload, using data from legacy operational systems and storing historical data to provision traditional business intelligence and analytics results.

But those operational systems can also populate the big data environment when they're needed for computation-rich processing or for raw data exploration. A company can steer the workload to the right platform based on what that platform was designed to do.

Figure 5: Big Data and Data Warehouse Coexistence

This minimizes disruption to existing analytics functions while at the same time accelerating new or strategic business processes that might benefit from increased speed. Figure 5 shows that the data warehouse can serve as a data source into the big data environment. Likewise, Hadoop can consolidate key data output that can populate the data warehouse for subsequent analytics.

> *"Sears is investing in real-time data acquisition and integration as it happens," says Sears' Oliver Ratzesberger. "We're bringing in open source solutions and changing our applications architecture. No more ETL…We're creating a framework that, over time, any application can leverage."*

By the end of 2013, there will be more mobile devices than people on the planet.[x] Harnessing data from a range of new technology sources gives companies a richer understanding of consumer behaviors and preferences—irrespective of whether those consumers are existing or future customers. Big data technologies not only scale to larger data volumes more cost effectively, they support a range of new data and device types. The flexibility of these technologies is only as limited as the organization's vision.

Integrating Big Data Technologies

An international financial services firm initially acquired a big data infrastructure to exploit faster processing power. But in every case, analytics is the next frontier. Managers we talked to are building out their big data roadmaps to solve a combination of both operational and analytical needs, many of them still unforeseen.

> *"The opportunities for cross-organizational analytics are huge," the Executive in charge of big data told us. "But when the firm's executives started discussing big data, the value-add was still esoteric. So we started instead by focusing on process efficiencies. We have 60 terabytes of what we consider to be analytics data sets, and we use compiled, multi-threaded code…and do periodic refreshes. We're past some of the challenges associated with 'fail fast' and are tapping into all the advantages of Hadoop."*

When determining the components of their big data environment, the executives we talked with asked some key questions, including:

1: What's the initial problem set we think new big data technologies can help us with?

2: What existing technologies will play a role?

3: Do we have the right skills in place to develop or customize big data solutions to fit our needs?

4: Do these new solutions need to "talk to" our incumbent platforms? Will we need to enable that? Are there open source projects that can give us a head start?

5: It's not practical for us to acquire all the big data-enabling technologies we need in one fell swoop. Assuming we can establish acquisition tiers for key big data solutions, what are the corresponding budget tiers?

By circumscribing a specific set of business problems, companies considering big data can be more specific about the corresponding functional capabilities and the big data projects or service providers that can help address them. This approach can inform both the acquisition of new big data technologies and the re-architecting of existing ones to fit into the brave new world of big data.

6. Big Data's Value Proposition

As we engaged big company executives in conversations about big data for this report, they all agreed that they considered big data to be an evolutionary set of capabilities that would have new and sometimes unanticipated uses over time. But every one of these executives conceded that they couldn't afford to make big data a mere academic exercise. It needed to drive value, and better sooner than later.

Return on Investment

Very few companies have taken the steps to rigorously quantify the return on investment for their big data efforts. The reality is that the proof points about big data often transcend hard dollars represented by cost savings or revenue generation. This suggests that senior management is betting on big data for the long-term, a point corroborated by several of the executives we interviewed.

But initial comparisons of big data ROI are more than promising. In 2011, Wikibon, an open source knowledge sharing community, published a case study[xi] that compared the financial return of two analytics environments. The first environment was a high-speed data warehouse appliance employing traditional ETL and data provisioning processes. The second environment was big data running on a newer big data technology using massively-parallel (MPP) hardware.

As Figure 6 shows, the project favored the MPP big data environment across a range of measures, including accelerating time-to-value (it showed value almost immediately after installation), cumulative cash flow, and internal rate of return.

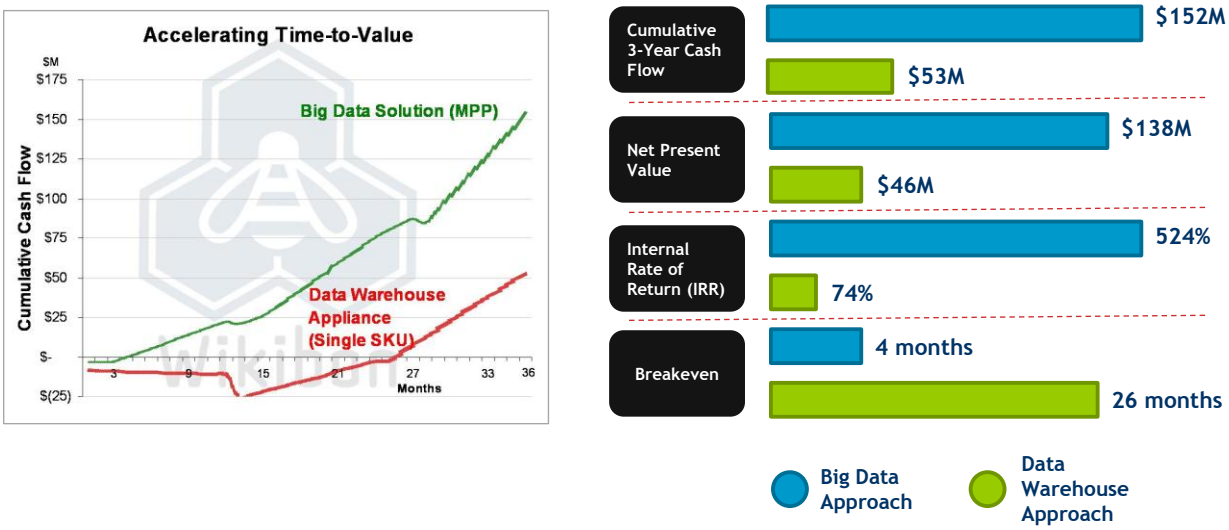

Figure 6: ROI Comparison for Big Data

(The findings became part of a larger discussion in the Wikibon community, which asked the question, "Is your data warehouse a dinosaur?") The study's conclusion was not that the data warehouse was becoming obsolete, but rather—as we outlined in Section 4—traditional data warehouses would end up peacefully coexisting with emerging big data solutions, each playing a specialized role in a company's analytics ecosystem.

The truth is that executives at big data early-adopter companies don't talk about how much money they're saving or making. Those we interviewed described two approaches for realizing return on big data. One approach was enabling new business capabilities for the first time. The other was doing what they were already doing, only cheaper, faster, or better than before.

Automating Existing Processes

Whether it's the need to do a proof-of-concept, explore preliminary data, or convince executives to invest, many companies have to prove the value of big data technologies as a first step to broader big data delivery. This often means delivering cost efficiencies or economies of scale within existing business paradigms.

Most of the executives we interviewed introduced big data technologies through an initial proof-of-concept approach to illustrate the high performance, lower cost of ownership, scale, and advanced business capabilities of big data solutions by applying them to current, often cumbersome, business processes.

Other companies see the promise of big data to bring together disparate platform and processing functions that were previously silo-ed. Our interviewees spoke aspirationally of the ability to combine data reporting, analytics, exploration, protection, and recovery functions on a single big data platform, thereby eliminating the need for complicated programming and specialized skills to tie legacy systems together.

Big Data at Sears

When it comes to the adoption of information technology, Sears was years ahead of most retailers, implementing an enterprise data warehouse in the 1980s while most retailers were still relying on manually-updated spreadsheets to examine their sales numbers. These days the company is using big data technologies to accelerate the integration of petabytes of customer, product, sales, and campaign data in order to understand increase marketing returns and bring customers back into its stores. The retailer uses Hadoop to not only store but process data transformations and integrate heterogeneous data more quickly and efficiently than ever.

"We're investing in real-time data acquisition as it happens," says Oliver Ratzesberger, Vice President of Information Analytics and Innovation at Sears Holdings. "No more ETL. Big data technologies make it easy to eliminate sources of latency that have built up over a period of time."

The company is now leveraging open source projects Apache Kafka and Storm to enable real-time processing. "Our goal is to be able to measure what's just happened."

The company's CTO, Phil Shelley, has cited big data's capability to decrease the release of a set of complex marketing campaigns from eight weeks to one week—and the improvements are still being realized. Faster and more targeted campaigns are just the tip of the iceberg for the retailer, which recently launched a subsidiary, MetaScale, to provide non-retailers with big data services in the cloud.

The good news about applying new technology to existing problems is that opportunities for improvement are already well-understood, and thus consensus is more easily achieved. "Fixing what's known to be slow or broken gets more support from my CEO than promoting new technologies out-of-the-box," explained a banking vice president. "He doesn't care whether our competitors are using big data. He cares that they could be gaining market share by making faster decisions."

Moreover, it's easier to measure new process improvements against traditional methods, so quantifying faster product time-to-market, higher return on marketing investment, or fewer patient readmissions makes quantifying return on investment is that much easier.

But even executives advocating large-scale change have their eye on the shiny, new capabilities promised by big data analytics.

Delivering the New

One of most positive outcomes of the big data trend is the way it has captured the attention of senior managers like no other technology trend before it. Suddenly, C-level executives are funding headcount for big data projects, and using the phrase "data as an asset" in board meetings.

New applications for big data are often industry-specific. Think telematics data for auto insurers, vital signs in health care, or RFID tags in manufacturing. All of this data is difficult to capture and ingest, let alone use in a meaningful way. A recent survey found that the majority of respondents, 41 percent, reported not having a strategy for big data. The next highest percentage of respondents reported "Operations/Processing" as being the area of focus for big data projects.[xii]

Clearly, most companies still haven't transcended their initial projects to articulate the full business potential of big data. It's still early days. And fundamental questions persist: Is big data best consumed by humans or machines? Is our most important data about our customers, or our operations? Will new data drive new insights, or simply confirm existing hypotheses? Most big companies are launching their big data projects starting with automation of existing processes and hoping to drive more strategic value. In most cases, that value is in the eye of the beholder.

Beyond Customers: Big Data's Infinite Promise

Go to any industry conference or read a vendor brochure, and it seems as if everyone's talking about big data driving new customer insights. Yes, you can troll through years' worth of customer data and quickly see which high-value customers have a propensity to churn. You can calculate lifetime customer value. You can look at purchase patterns to see what a business customer might buy next. And you can

Big Data at GE

"It wasn't just Jeff,"—as in Immelt, GE's CEO—"who saw the potential of this," said Bill Ruh, Vice President and Corporate Officer for GE's Global Software Center. "There were a number of senior executives who saw the benefits from the marriage of machines and analytics. For instance, the trend of customers wanting to optimize inspection, maintenance and repair processes of our machines. Or even the potential for our machines to communicate with one another or with operators to enable intelligent decisioning. These type of instrumentation opportunities are present in locomotives, power plants and industrial facilities, so needs were mushrooming all around the company."

Ruh highlights GE's industrial business as a prime target for big data, referencing the health of blades on the jet engines the company manufactures. "Our sensors collect signals on the health of blades on a gas turbine engine to show things like 'stress cracks.' The blade monitor can generate 500 gigabytes per day—and that's only one sensor on one turbine. There are 12,000 gas turbines in our fleet." The value in integrating all the sensor data onto a big data platform can reveal patterns on when blades break, allowing GE to tune its manufacturing and repair process before a break occurs.

"Most companies aren't ready for real-time data like this," Ruh says. "They know when to make the decisions. But we'll provide insight on how to operate the jet engine more efficiently, or when to adjust the gas turbine. The way we'll get to the proverbial 'Power of One' [the value of a one percent improvement in turbine efficiency, which totals in the multiple billions] is in our ability to operationalize this."

develop micro-segments and corresponding microsites for key customer clusters, and communicate with them in increasingly relevant ways.

Strictly speaking, you don't need big data for any of this. This is customer analytics—sometimes called "analytical CRM," sometimes called "business intelligence." And companies across industries and market segments have been doing it long before big data was a buzzword and statisticians were sexy.

It's the consumer's "digital footprint" from online purchases, in-store kiosk interactions, ATM transactions, and social media commentary that's resulting in part of the big data explosion. These interactions make behavioral analytics and targeting that much richer—and more interesting to companies, their advertisers, and third-party data providers.

The consumer who searches the web for camping gear, top-of-the-line fly fishing rods, and family vacations packages in Montana may be a better candidate for a zero-interest loan on a four-wheel drive truck than the shopper comparing parkas with faux-fur trim. But depending on other interactions or interests revealed through richer behavior and preference data, either might be a good candidate for an eco-friendly volunteer vacation. The examples of the power of big data analytics to drive customer loyalty are endless.

7. The Rise of Analytics 3.0

In order to understand the role of big data in big companies, it's important to understand the historical context for analytics and the brief history of big data. Analytics, of course, are not a new idea. The tools have been used in business since the mid-1950s. To be sure, there has been a recent explosion of interest in the topic, but for the first half-century of activity, the way analytics were pursued in most organizations didn't change that much. We'll call the initial era *Analytics 1.0*. This period, which stretched 55 years from 1954 (when UPS initiated the first corporate analytics group) to about 2009, was characterized by the following attributes:

- Data sources were relatively small and structured, and came from internal sources;
- Data had to be stored in enterprise warehouses or marts before analysis;
- The great majority of analytical activity was descriptive analytics, or reporting;
- Creating analytical models was a "batch" process often requiring several months;
- Quantitative analysts were segregated from business people and decisions in "back rooms";
- Very few organizations "competed on analytics"—for most, analytics were marginal to their strategy.

In the period from 2005 to 2012, the world began to take notice of big data, and we'll have to call that the beginning of *Analytics 2.0*. The era began with the exploitation of online data in Internet-based firms like Google, Yahoo, and eBay. Big data and analytics not only informed internal decisions, but also formed the basis for customer-facing products and processes. However, large companies often confined their analytical efforts to basic information domains like customer or product that were highly-structured and rarely integrated with other data.

Big data analytics as a standalone entity in Analytics 2.0 were quite different from the 1.0 era in many ways. Data was often externally-sourced, and as the big data term suggests, was either very large or unstructured. The fast flow of data meant that it had to be stored and processed rapidly, often with massively parallel servers running Hadoop. The overall speed of analysis was much faster. Visual analytics—a form of descriptive analytics—often crowded out predictive and prescriptive techniques. The new generation of quantitative analysts was called "data scientists," and many were not content with working in the back room; they wanted to work on new product offerings and to help shape the business.

Big data, of course, is still a popular concept, and one might think that we're still in the 2.0 period. However, there is considerable evidence that large organizations are entering the *Analytics 3.0* era. It's an environment that combines the best of 1.0 and 2.0—a blend of big data and traditional analytics that yields insights and offerings with speed and impact. Although it's early days for this new model, the traits of Analytics 3.0 are already becoming apparent. The most important trait is that not only online firms, but virtually any type of firm in any industry, can participate in the data-driven economy. Banks, industrial manufacturers, health care providers, retailers—any company in any industry that is willing to exploit the possibilities—can all develop data-based offerings for customers, as well as supporting internal decisions with big data.

> *"From the beginning of our Science function at AIG, our focus was on both traditional analytics and big data. We make use of structured and unstructured data, open source and traditional analytics tools. We're working on traditional insurance analytics issues like pricing optimization, and some exotic big data problems in collaboration with MIT. It was and will continue to be an integrated approach."—Murli Buluswar, Chief Science Officer, AIG*

Other attributes of Analytics 3.0 organizations are described below:

Multiple Data Types, Often Combined

Organizations are combining large and small volumes of data, internal and external sources, and structured and unstructured formats to yield new insights in predictive and prescriptive models. Often the increased number of data sources is incremental, rather than a revolutionary advance in capability. At Schneider National, for example, a large trucking firm, the company is increasingly adding data from new sensors—monitoring fuel levels, container location and capacity, driver behavior, and other key indicators—to its logistical optimization algorithms. The goal is to improve the efficiency of the company's route network, to lower the cost of fuel, and to decrease the risk of accidents (see the "Big Data at Schneider National" case study).

A New Set of Integration Options

Since the 1970s, IT organizations have largely organized operational data in the relational database format. Since the 1980s, they have employed data warehouses with copies of operational data as the basis for analysis. Now, however, there are a new set of options from which to choose: Database appliances, Hadoop clusters, SQL-to-Hadoop environments, etc. The complexity and number of choices

that IT architects have to make about data management have expanded considerably, and almost every organization will end up with a hybrid environment. The old formats haven't gone away, but new processes need to be developed by which data and the focal point for analysis will move across staging, evaluation, exploration, and production applications.

Technologies and Methods Are Much Faster

Big data technologies include a variety of hardware/software architectures, including clustered parallel servers using Hadoop/MapReduce, in-memory analytics, in-database processing, and so forth. All of these technologies are considerably faster than previous generations of technology for data management and analysis. Analyses that might have taken hours or days in the past can be done in seconds. To complement the faster technologies, new "agile" analytical methods and machine learning techniques are being employed that produce insights at a much faster rate. Like agile system development, these methods involve frequent delivery of partial outputs to project stakeholders; as with the best data scientists' work, there is an ongoing sense of urgency. The challenge is adapting operational and decision processes to take advantage of what the new technologies and methods can bring forth.

Big Data at Schneider National

Schneider National, one of North America's largest truckload, logistics and intermodal services providers, has been pursuing various forms of analytical optimization for a couple of decades. What has changed in Schneider's business over the past several years is the availability of low-cost sensors for its trucks, trailers and intermodal containers. The sensors monitor location, driving behaviors, fuel levels and whether a trailer/container is loaded or empty. Schneider has been transitioning to a new technology platform over the last five years, but leaders there don't draw a bright line between big data and more traditional data types. However, the quality of the optimized decisions it makes with the sensor data – dispatching of trucks and containers, for example – is improving substantially, and the company's use of prescriptive analytics is changing job roles and relationships.

New sensors are constantly becoming available. For example, fuel-level sensors, which Schneider is beginning to implement, allow better fueling optimization, i.e., identifying the optimal location at which a driver should stop for fuel based on how much is left in the tank, the truck's destination and fuel prices along the way. In the past, drivers have entered the data manually, but sensor data is both more accurate and free of bias.

Safety is a core value at Schneider. Driving sensors are triggering safety discussions between drivers and their leaders. Hard braking in a truck, for example, is captured by sensors and relayed to headquarters. This data is tracked in dashboard-based safety metrics and initiates a review between the driver and his/her leader. Schneider is piloting a process where the sensor data, along with other factors, goes into a model that predicts which drivers may be at greater risk of a safety incident. The use of predictive analytics produces a score that initiates a pre-emptive conversation with the driver and leads to less safety-related incidents.

Integrated and Embedded

Consistent with the increased speed of analytics and data processing, models in Analytics 3.0 are often being embedded into operational and decision processes, dramatically increasing their speed and impact. Some are embedded into fully automated systems based on scoring algorithms or analytics-based rules. Some are built into consumer-oriented products and features. In any case, embedding the analytics into systems and processes not only means greater speed, but also makes it more difficult for decision-makers to avoid using analytics—usually a good thing.

New and Hybrid Technology Environments

It's clear that the Analytics 3.0 environment involves new technology architectures, but it's a hybrid of well-understood and emerging tools. The existing technology environment for large organizations is not being disbanded; some commented that they still make effective use of relational databases on IBM mainframes. However, there is a greater use of big data technologies like Hadoop on commodity server clusters; cloud technologies (private and public), and open-source software. The most notable changes are attempts to eliminate the ETL (extract, transform, and load) step before data can be assessed and analyzed. This objective is being addressed through real-time messaging and computation tools such as Apache Kafka and Storm. A related approach being explored is a new discovery platform layer of technology for data exploration. Enterprise data warehouses were initially intended for exploration and analysis, but they have become production data repositories for many organizations, and getting data into them requires expensive and time-consuming ETL work. Hence the need for a new layer that facilitates data discovery.

Data Science/Analytics/IT Teams

Data scientists often are able to run the whole show—or at least have a lot of independence—in online firms and big data startups. In more conventional large firms, however, they have to collaborate with a variety of other players. In many cases the "data scientists" in large firms may be conventional quantitative analysts who are forced to spend a bit more time than they like on data management activities (which is hardly a new phenomenon). And the data hackers who excel at extracting and structuring data are working with conventional quantitative analysts who excel at modeling it. Both groups have to work with IT, who supplies the big data and analytical infrastructure, provisions the "sandboxes" in which they can explore data, and who turns exploratory analyses into production capabilities. Together the combined teams are doing whatever is necessary to get the analytical job done, and there is often a lot of overlap across roles.

Chief Analytics Officers

It wouldn't make sense for companies to have multiple leaders for different types of data, so they are beginning to create "Chief Analytics Officer" roles or equivalent titles to oversee the building of analytical capabilities. For example, AIG brought in long-term analytics leader Murli Buluswar to be "Chief Science Officer" at the company—perhaps the only official "C-level" analytics executive in any large firm. He oversees a variety of analytical projects and groups, involving both big data and small. His staff includes data scientists and conventional quantitative analysts. The group works on traditional

insurance problems (e.g., analytical pricing optimization), and is collaborating with MIT researchers on big data projects. Buluswar is representative of this new type of leader under Analytics 3.0.

The Rise of Prescriptive Analytics

There have always been three types of analytics: descriptive, which report on the past; predictive, which use models based on past data to predict the future; and prescriptive, which use models to specify optimal behaviors and actions. Analytics 3.0 includes all types, but there is an increased emphasis on prescriptive analytics. These models involve large-scale testing and optimization. They are a means of embedding analytics into key processes and employee behaviors. They provide a high level of operational benefits for organizations, but they place a premium on high-quality planning and execution. Prescriptive analytics also can change the roles of front-line employees and their relationships with supervisors, as at Schneider National.

Summary

Even though it hasn't been long since the advent of big data, these attributes add up to a new era. It is clear from our research that large organizations across industries are joining the data economy. They are not keeping traditional analytics and big data separate, but are combining them to form a new synthesis. Some aspects of Analytics 3.0 will no doubt continue to emerge, but organizations need to begin transitioning now to the new model. It means change in skills, leadership, organizational structures, technologies, and architectures. It is perhaps the most sweeping change in what we do to get value from data since the 1980s.

It's important to remember that the primary value from big data comes not from the data in its raw form, but from the processing and analysis of it and the insights, products, and services that emerge from analysis. The sweeping changes in big data technologies and management approaches need to be accompanied by similarly dramatic shifts in how data supports decisions and product/service innovation. There is little doubt that analytics can transform organizations, and the firms that lead the 3.0 charge will seize the most value.

This independent research study was conducted by Tom Davenport and Jill Dyché, and was sponsored by SAS. For more information on this topic, or to participate in this ongoing research, please contact Tom Davenport at thdavenport@gmail.com or Jill Dyché at Jill.Dyche@sas.com. To learn more about SAS visit www.sas.com. To learn more about the International Institute for Analytics visit iianalytics.com

About the Authors:

Thomas H. Davenport is a Visiting Professor at Harvard Business School, a distinguished professor at Babson College, a Senior Advisor to Deloitte Analytics, and co-founder and research director of the International Institute for Analytics. He has co-authored or edited four books on business analytics, including the new book *Keeping Up with the Quants: Your Guide to Understanding and Using Analytics*.

Jill Dyché is Vice President of Best Practices at SAS, and the author of three books on the business value of technology. Her work has been featured in major publications such as *Computerworld*, *the Wall Street Journal*, and *Newsweek.com*, and she blogs on technology trends for Harvard Business Review. Jill was the co-founder of Baseline Consulting, which was acquired by SAS in 2011.

---

[i] NewVantage Partners, "Big Data Executive Survey: Themes and Trends," 2012.
[ii] Peter Evans and Marco Annunziata, "Industrial Internet: Pushing the Boundaries of Minds and Machines," GE report, Nov. 26, 2012. *www.ge.com/docs/chapters/Industrial_Internet.pdf*
[iii] The M Group's use of big data is described in Joel Schectman, "Ad Firm Finds Way to Cut Big Data Costs," Wall Street Journal CIO Journal website, February 8, 2013, http://blogs.wsj.com/cio/2013/02/08/ad-firm-finds-way-to-cut-big-data-costs/
[iv] Kerem Tomak, in "Two Expert Perspectives on High-Performance Analytics," Intelligence Quarterly (a SAS publication), 2nd quarter 2012, p. 6.
[v] Tom Vanderbilt, "Let the Robot Drive: The Autonomous Car of the Future Is Here," *Wired*, January 20, 2012, http://www.wired.com/magazine/2012/01/ff_autonomouscars/
[vi] Andrew Leonard, "How Netflix Is Turning Viewers into Puppets," February 1, 2013, http://www.salon.com/2013/02/01/how_netflix_is_turning_viewers_into_puppets/
[vii] Open source solutions are known as "projects" because they are developed jointly by a community of contributors. Thus, they represent a collection of diverse and often far-flung activities that, when unified, comprise a holistic solution. Because they are built by a community of developers who are typically unpaid for their work (many accept donations), these projects are often free-of-charge to individuals or companies who contribute additional functionality or guidance to the community. This is the opposite of proprietary software solutions, which are pre-packaged as products with finite release schedules and more rigid pricing models. By their very nature, open source projects are ongoing, until the community stops using the software and/orthe members of the developer community stop contributing to them.
[viii] SAS 2013 Big Data Survey, page 1: http://www.sas.com/resources/whitepaper/wp_58466.pdf.
[ix] "Big Data: The Next Frontier for Innovation, Competition, and Creativity," McKinsey Global Institute, 2011.
[x] According to a report published by Cisco Systems, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update," February 6, 2013. http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html
[xi] For the complete story on this study, see http://wikibon.org/wiki/v/Financial_Comparison_of_Big_Data_MPP_Solution_and_Data_Warehouse_Appliance.
[xii] SAS 2013 Big Data Survey, page 4: http://www.sas.com/resources/whitepaper/wp_58466.pdf.