Research Report

The Human Side of Big Data and High-Performance Analytics

August 2012

Authored by:

Thomas H. Davenport

iianalytics.com

Executive Summary

In order to succeed with big data and high-performance analytics, organizations require not only new technologies, but also a new set of human capabilities. A key component of these capabilities are data scientists—hybrids of analytical and computational skills, typically with science backgrounds. Data scientists are motivated not just to support internal decisions with analytics, but to create new products and processes for customers. In addition, big data and high-performance analytics require new approaches to deciding and acting on the part of executives and decision-makers.

The concept of "big data" burst onto the technology and business scene in 2010, and since then has excited many executives with its potential to transform businesses and organizations. The concept refers to data that is either too voluminous, too unstructured, or from too many diverse sources to be managed and analyzed through traditional means. The definition is clearly a relative one that will change over time. At the moment, however, "Too voluminous" typically means databases or data flows in petabytes (1000 terabytes); Google, for example, processes about 24 petabytes per day of data. "Too unstructured" generally means that the data doesn't initially come in numeric form, and isn't easily put into the traditional rows and columns of conventional databases.

Examples of big data include a massive amount of online information, including clickstream data from the web, and social media content (tweets, blogs, wall postings, etc.). Big data also incorporates video data from retail and crime/intelligence environments, location data from mobile devices, and data from the rendering of video footage. It includes voice data from call centers or intelligence interventions. In the life sciences, it includes genomic and proteomic data from biological research and medicine.

Many IT vendors and solutions providers, and some of their customers, are treating the term as just another buzzword for managing and analyzing data to better understand the business. But there is more than vendor hype; there are considerable business benefits from being able to analyze big data on a consistent basis. And given the pervasiveness of devices that generate data—from smartphones to RFID chips to sensors in manufacturing technologies and manufactured products—we can expect big data in almost every industry.

Companies that excel at big data will be able to employ these technologies with business benefit, and to harness the "Internet of things." They will produce new products and services based on data. They will ultimately be able to understand their business environment at the

most granular level, and adapt to it rapidly. They'll be able to differentiate commodity products and services by incorporating monitoring and analysis of usage patterns. And in the life sciences, of course, effective use of big data can yield cures to the most threatening of diseases.

Companies with more traditional types of structured data can also benefit from recent technology developments. Advances in parallel computing, in-database analytical processing, and in-memory software technologies now make it possible to dramatically accelerate the pace with which analytical insights can be delivered. Often called "high-performance analytics" (HPA), these tools can reduce the cycle time for complex and large-scale analytical calculations from hours or even days to minutes or seconds. The department store chain Macy's, for example, 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. HPA obviously makes it possible to reprice items much more frequently to adapt to changing conditions in the retail marketplace.

But prospering from big data is not simply a matter of employing new technologies. To achieve the benefits from big data and high-performance analytics, firms will need to make some adjustments to their capabilities, even if they are already experienced users of analytics. Big data and high-performance analytics environments are clearly different from traditional data analysis environments in many ways. In this white paper, however, the focus is on the human-related differences. A key focus is on the "data scientists" who do this sort of work, and the implications for executives and managers will also be considered. See the sidebar "About the Research" for more details on the research approach employed.

Sidebar: About the Research

In order to learn more about data scientists and their activities, I interviewed (primarily by telephone) 30 people in the first half of 2012 who described themselves as data scientists, and several more managers to whom data scientists reported. In all but a couple of cases, it was clear that the individuals interviewed had data scientist-oriented backgrounds and were actually performing data science activities. The data scientists came from a wide variety of industries, including online data and services (8 of 30, or 27%), IT vendors (20%), professional or IT services (13%), health care (13%), manufacturing (7%), media (7%), and several industries with only one respondent.

I asked about a variety of topics, including their backgrounds, motivations for entering work in big data, current activities, reporting and working relationships, and views on what makes for effective data science teams.

The Rise of Data Scientists

While there has always been a need for analytical professionals to support an organization's analytical capabilities, with big data the requirements for support personnel are a bit different. The fact that dealing with the data itself—obtaining, extracting, manipulating, and structuring it—is an important prerequisite to doing any analysis means that the people who work with big data need substantial IT skills.

The role of analytical professional in big data environments has been christened as the "data scientist." Data scientists understand analytics, but they also are quite skilled at exploring and exploiting information technology. Data scientists work in a variety of organizations, from big data startups (37% of sample interviewed) to large, established companies like GE, Intuit, and EMC (63%). GE, for example, expects to hire over 400 computer and data scientists at its new Global Software and Analytics Center in the San Francisco Bay Area. While some of the GE data scientists will work on traditionally data-intensive problems in financial services, logistics, and health care, a significant portion of their focus will be on big data for industrial products, such as locomotives, turbines, jet engines, and large energy generation facilities.

Data scientists have somewhat different roles from traditional quantitative analysts. Whereas traditional analysts typically use analytics on internally generated data to support internal decisions, the focus of many data scientists is on customer-facing products and processes, where they help to generate products, features, and value-adding services. At the business social network LinkedIn, for example, data scientists developed the "People You May Know" and "Jobs You May Be Interested In" features of the site, among others. GE is already using data science to optimize the service contracts and maintenance intervals for industrial products. Google, of course, uses its roughly 600 data scientists to refine its core search and ad-serving algorithms. Zynga uses data scientists to target games and game-related products to customers. The testing firm Kaplan uses its data scientists to begin advising customers on effective learning and test-preparation strategies. In health care big data firms, data scientists try to discover the most effective treatments for different diseases.

Given this focus, data scientists are most likely to be in product development or marketing organizations (just over half of my sample) Some also work for Chief Technology Officer (CTO) organizations. Some firms have established new positions to which data scientists report, including a Senior Vice President of Big Data, Social Design, and Marketing (at Intuit), and a Chief Digital Officer (at Kaplan). Data scientists are not likely to work closely with corporate IT organizations or Chief Information Officers, though stronger relationships with those functions

are sometimes found in health care big data firms. Several did mention, however, that they relied on traditional "business intelligence" functions for help with reporting of the results from big data analyses.

Those who report to CTOs are likely to work on tools that make data science easier and more productive. Because the tools for managing large volumes of unstructured data are still evolving rapidly (the widely used Hadoop framework for distributed file system processing only went into production at Yahoo in 2008, for example), some data scientists are focused on creating and improving tools. Yahoo thus far hasn't used data as a product, but rather built an infrastructure for the rest of the world to use. This has also been a primary activity, for example, of Facebook's data team, which has created the language Hive for programming Hadoop projects.

Some firms, such as Yahoo, Facebook, and Microsoft, also support academic-style research and publication by their data scientists. This may be useful for retention of the data scientists, although the business value of the activity is less clear.

Data scientists who focus on HPA applications don't necessarily need to understand how to process unstructured data, but they do need to understand how analytical work can be divided across multiple parallel servers. They may also serve as process consultants, helping their organizations to speed up business processes based on the dramatic reductions in time required to produce extensive analytical results. They should be able to explore a variety of ways to use the extra time from HPA to refine their models. In addition, as I will discuss in greater detail below, they need to try to accelerate decision speeds to match the much faster cycle times of data analysis.

Data Scientist Skills

Data scientists require technical, business, analytical, and relationship skills. From a technical standpoint, many have advanced computer science degrees, or advanced degrees (often Ph.D.s—57% of my sample) in fields such as physics, biology, or social sciences that require a lot of computer work. 90% of the data scientists I interviewed had at least one degree in a scientific or technical field. Almost all had strong computational skills. They're not just programmers; many refer to data scientists' computational skills as "hacking" technology—bending it to do their bidding in unusual ways.

LinkedIn, the social networking site targeted at business professionals, has a substantial data science team, and has had considerable success developing data-driven products and features. A job description for the data scientist role at LinkedIn describes the desired technology traits:

> Are you someone who solves hard problems by creatively obtaining, scrubbing, exploring, modeling and interpreting big data? Do you know enough about information retrieval, machine learning, and statistics to be dangerous? Are you a hands-on implementer, ready to learn new languages and technologies to turn your ideas into solutions used by tens of millions of people around the world?

As the job description suggests, data scientists tend to address a few key technologies. They include:

- Hadoop, MapReduce, and the related ecosystem of distributed file system tools;
- Programming languages such as Python, Java, Pig and Hive;
- Machine learning;
- Nontraditional database tools such as Vertica and MongoDB;
- Natural language processing;
- Statistical tools.

Overall, data scientists—particularly those in startup firms—seem to have a strong preference for open-source tools and technologies.

In addition to these technical skills, data scientists also need the attributes previously necessary for analytical professionals. These include mathematical and statistical skills, business acumen, and the ability to communicate effectively with customers, product managers, and decision-makers. Several data scientists commented that relationships are important to success in the field. This was also true with traditional quantitative analysts, but the relationships there were with internal business decision-makers, rather than product and process managers.

Of course, the combination of these skills is difficult to find in one person. Some companies, reflecting this problem, have created data science teams that together embody this collection of skills. Each individual member of the team may have only some of the necessary skills, but assuming they work closely together, they can do all the necessary activities.

Recruiting or Creating Data Scientists

Adding the hard-core technical requirement to traditional analytical skills makes it even more difficult to find such individuals. At a recent gathering in Silicon Valley on big data, the consensus of the attending experts was that finding qualified data scientists in sufficient numbers will be the greatest constraint on the field. And in a recent Economist Intelligence Unit survey of 600 global executives, 54% of North American respondents said finding the right people with the right skills is the No. 1 obstacle to launching a successful big data project.

Where can an organization find data scientists? There are few if any academic programs in the area, although several are being designed now. Some existing master's degree programs in analytics, such as that at North Carolina State, are including some big data training in their curricula (such as Hadoop programming and dealing with unstructured data). Most organizations, however, must recruit and hire individuals from other backgrounds who have skills related to data science.

For example, George Roumeliotis, the head of a data science team at Intuit in Silicon Valley (and himself a Ph.D. in astrophysics), doesn't hire purely on the basis of statistical or analytical capabilities; instead, he needs people who can develop prototypes in a mainstream programming language such as Java. He says he has given up trying to recruit anyone with industry experience—they just don't exist—so he primarily recruits directly out of schools (Ph.D. programs). Roumeliotis seeks both a skillset—a solid foundation in math, statistics, probability, and computer science—and a mindset—a feel for business issues and an empathy for customers. Then, he says, he builds on that with a mixture of on-the-job training and an occasional course in a particular technology.

Given the difficulty of finding and keeping data scientists, one would think that a good strategy would involve hiring them as consultants. Yet most firms that are aggressively pursuing big data projects seem to want to employ their own data scientists (perhaps because they are worried about turning over their important data assets to outside firms), and most consulting firms have yet to assemble them in large numbers. Firms such as Accenture, Deloitte, and IBM Global Services do have data scientists on staff (some call them "management scientists" or "decision scientists"), but they are only in the early stages of leading big data projects for large-firm clients. Thus far they are primarily applying their skills to more conventional quantitative analysis problems. As client demand for big data work builds, they will no doubt offer more data scientists-for-hire.

There are a variety of other approaches in use to develop and hire data scientists. EMC, for example, has determined that the availability of data scientists will be an important gating factor in its own big data efforts and those of its customers. So it has created a "Data Science and Data Analytics" training program for its employees and customers. EMC has already begun to put graduates of the program to work on internal big data projects, and has also made the course materials available to universities.

One data scientist has come up with a creative approach to training new data scientists. The Insight Data Science Fellows Program, started by Jake Klamka (a high energy physicist by background) takes scientists for 6 weeks and teaches them the skills to be a data scientist. The program includes mentoring by local companies with big data challenges (e.g., Facebook, Twitter, Google, LinkedIn, etc). "I originally was aiming for 10 Fellows. I had over 200 applicants and accepted 30 of them, "says Klamka. He goes on, "The demand from companies has been phenomenal; they just can't get this kind of high quality talent."

Venture capital firms are also entering the data science game. In order to help the demand by companies in their portfolio, Greylock Partners, an early stage venture firm that has backed companies like Facebook, LinkedIn, Palo Alto Networks, and Workday, has a built a recruiting team that focuses in part on data scientists. Dan Portillo, who leads the team, says, "The demand of data scientists is at an all-time high from our later stage companies. Once they have data, they really need people who can manage it and find insights in it. The traditional backgrounds of people you saw 10-15 years ago, just don't cut it these days."

Once they are hired or created, companies may also face issues in retaining data scientists. Several of those I interviewed had changed jobs several times in the past year. One commented, "After about a year it often becomes clear that there is nothing left for me to do." Another noted, "Data scientists receive lots of job offers—sometimes I get two or three calls a week from headhunters. It's not surprising that with so much opportunity there is a lot of movement." In order to hold on to the data scientists in your organization, you need to offer them a combination of autonomy (to be able to make an impact with their work), interesting and useful data, and a state-of-the-art technical environment—all in addition to a lucrative compensation package.

What Motivates Data Scientists?

If you're interested in recruiting and retaining data scientists for big data work, it's important to know what motivates them. In my interviews with data scientists, the same motivational issues

appeared frequently. Data scientists want to use data to have a substantial impact on the world. They view this as a unique period in history in which there are huge datasets and very powerful tools. As Amy Heinike, a prominent data scientist at the startup Quid in San Francisco put it:

> If you have access to the data and the tools, you can already find out some really cool stuff, but we are just scratching the surface. What inspires me is the opportunity to create something really interesting. Could something be important, have impact, or reach a lot of people? I am also interested in how to include data scientists within a diverse engineering team and company, and in combining the diverse skills that make up an effective data science team. So when I evaluate an opportunity, I look for a rich dataset that the company has to work with, or an important question for which we might be able to find data. And I want to make sure that there are the resources and senior management support available to support the data science function.

Another data scientist noted that he and his colleagues are motivated by—and highly attuned to—the attitudes of the founders of their company. He noted: "The key issue is how analytical are the founders—how ambitious and open-minded are they? That's ultimately the deciding factor in how analytical the organization becomes. If they're just oriented to the technology and the engineering, it won't happen."

Data scientists seem more motivated by the challenge and impact of the work than by money. One commented, "If we wanted to work with structured data, we'd be on Wall Street." Indeed, it seems likely that the "quants" who went to Wall Street in substantial numbers in the 1990s are the same types of people who become data scientists today.

Data scientists also see creating a product as far more motivating than simply advising a decision-maker. One described being a consultant as "the dead zone—all you get to do is tell someone else what the analyses say they should do." They view working on products or processes for customers as having much more potential impact. A new feature of an online social site, a new online game, even a new way of treating cancer—these product/process offerings are appealing to data scientists, and for good reason. Of course, to implement such offerings successfully requires that data scientists have both a high level of autonomy, and a close relationship with the leaders of product and process development organizations.

Problems with the Data Science Profession

In addition to the challenges of finding and retaining data scientists—and of motivating them sufficiently—there are some other potential difficulties with the role and the profession that may deter some organizations from employing them. Not surprisingly, data scientists are also expensive (whether primarily motivated by money or not), with many of those in startup organizations pulling down large options packages.

Another problem is that while a primary attribute of data scientists is the combination of technology-intensive "data wrangling" and analytics, there is often more of the former than the latter. Some of the data scientists interviewed suggest that "big data often equals small math." The amount of effort necessary to deal with large volumes of unstructured data sometimes means there is little time and resources left over for detailed statistical analysis. The primary analytical approach is descriptive analytics (also known as reporting); this is one reason why many data scientists are oriented to visual analytics, which typically don't work as well on predictive or prescriptive approaches.

Current data scientists also often face issues of relatively low productivity. The fact that new programs have to be written just to extract and structure data makes it labor-intensive and time-consuming to do data science work. One data scientist at GE works on extracting data from locomotive components; he complained that it took several months just to extract data from the alternator alone.

The next generation of data scientists will undoubtedly be more productive and will use tools that make common tasks much easier. Even then, however, there is likely to be considerable integration activity necessary to make use of big data.

A Different Approach to Deciding

Data scientists aren't the only human concern for big data and high-performance analytics; they also have an impact on decision-making by managers and executives. As noted above, big data is often used for purposes of product development rather than internal decision-making within organizations. However, when it does support decisions, the volume and velocity of big data is such that conventional, high-certitude approaches to decision-making are often not appropriate. By the time an organization has achieved a high level of certitude about the

insights and the implications, much more new data would have become available. Therefore, many organizations must adopt a more continuous, more indicative, less certain approach to analysis and decision-making.

Social media analytics, for example, are rarely definitive; instead, they provide fast-breaking trends on customer sentiments about products, brands, and companies. It might be useful to determine with certainty whether an hour's or a day's online content is correlated with sales changes, but by the time that analysis would be completed, much more new content would have arrived. It's important then, to have clear criteria for what decisions to make and what actions to take based on big data analyses—particularly in fast-changing domains like social analytics.

Sometimes it's important to admit that the data and analyses are not definitive. Even the United Nations—an organization typically not known for its agility—is getting in on this new approach to deciding. The UN's Global Pulse innovation lab has developed a big data-related tool called Hunchworks. The idea is that as data begin to reveal a trend or finding—say, for example, weather data suggesting a drought that could lead to famine in a part of Africa—the analyst would post the hunch and the data on which it was based, and others could weigh in with new analyses and data. One goal is to determine how likely the hunch is to be worthy of detailed analysis and action. But the idea that the UN would have a system for circulating data-driven hunches marks a major change in that organization's culture.

In a related but broader sense, organizations managing big data need to view data management, analysis, and decision-making in terms of "industrialized" flows and processes, rather than discrete stocks of data or events. Historically, data analysis required significant time and human intervention. Once data was identified, it was extracted and loaded into a data warehouse, and then analysts went to work. Typically the analysts would take significant time in massaging, analyzing, and interpreting the data. In many cases they would report it out in visual formats for better understanding by decision-makers.

However, the volume and velocity of big data means that organizations must develop continuous processes for gathering, analyzing, interpreting, and acting on data—at least for operational applications such as multi-channel customer "next best offers," identifying fraud in real time, and scoring medical patients for health risk. Much analysis, and at least some decision-making, must be automated or semi-automated.

As noted previously, big data analyses often involve reporting of data in visual formats. While there are increasing technological capabilities for displaying data in dashboards and visual

analytics, even visual analytics often require considerable human interpretation. The time and expense of such human interpretation may be difficult to justify in big data environments. As much as possible, humans should be eliminated from operational big data processes, or their involvement limited to initial development of rules and scoring algorithms, and dealing with exceptions. Of course, there will still be a role for detailed human analysis in research-oriented applications, but these contexts will typically involve slower analysis and decision processes.

In high-performance analytics (HPA) environments, the technology allows for much faster analysis of data. But if organizations are to receive value, they need to determine what to do with the additional time that has been freed up. Can they, for example, do more analyses to refine their models? One retail company was spending 5 hours to develop a single algorithm model each day for new customer acquisition. Using HPA approaches, the company reduced the model processing time to only 3 minutes, allowing a model to be iterated every 30 minutes or so while also using multiple modeling techniques. This improved the model lift from 1.6% to 2.5%--a seemingly small improvement, but one that could pay off dramatically across many customers. Other ways to refine models with HPA include using more data, incorporating more variables, and trying to fit more varieties of models through machine learning approaches.

Some other companies are attempting to improve the entire process in which the HPA work takes place. The key to this is accelerating decisions and actions to match the increased speed of analysis. Perhaps the freed-up time from HPA could be used to allow everyone to go on vacation, but this is unlikely to provide the ROI that many companies desire.

Summary and Future Directions

Just as traditional quantitative analysis on “small data” didn’t happen without professional and semi-professional analysts, big data can’t be analyzed without data scientists. Now that there is plenty of data available in most industries, and a substantial amount of technology is available to manipulate big data, all that is still required to make it useful is a qualified data scientist. Such a person can not only convert unstructured to structured data and perform quantitative analysis on it, but also help an organization think about what data sources to investigate, what customers really need in data and analysis requirements, and how best to incorporate big data-based products and services into an effective business model. Each of these roles is important, and data scientists are the single biggest constraint on the ability of organizations to capitalize on big data.

Of course, it's not easy to hire and retain data scientists today, and there will continue to be a shortage of them for several years. Eventually, university programs will arise to produce data scientists in larger numbers—not unlike the proliferation of masters-level programs in financial engineering for Wall Street quants. However, even traditional "small data" analysts are in short supply in many markets, so a constrained labor market for data scientists is something that organizations need to plan for and adapt to. The most successful big data organizations will be those that create or identify unique sources of data science talent.

We are just getting started in identifying what big data and high-performance analytics will do for decision-making and other organizational processes. Since much of big data work thus far is not primarily focused on internal decisions, it may be several years before we fully understand the implications of big data for decision-making. It's already clear, however, that simply standing in the big data river is not enough; organizations and individual managers must think carefully about converting big data insights into specific decisions and actions.

Many executives are excited about the potential of big data and high-performance analytics for their organizations. But they need to realize that putting big data to work requires a special breed of analyst. Even if an organization isn't quite ready to aggressively pursue big data opportunities yet, it's worth thinking now about how and when it will acquire the most scarce and valuable resource in big data—the data scientists.

About the Author and Sponsors

Thomas H. Davenport is a Visiting Professor at Harvard Business School, co-founder and research director of the International Institute for Analytics (IIA), and a Senior Advisor to Deloitte Analytics. This independent research study was conducted by Thomas Davenport and the IIA, and was co-sponsored by SAS Institute, Inc. and Greenplum, a division of EMC Corporation. To learn more about these companies, visit their respective websites at www.sas.com/big-data and www.greenplum.com. For more information on this topic or the research, please contact Tom Davenport at tdavenport@hbs.edu.