The Culture of Big Data

Mike Barlow

O'REILLY®

O'REILLY®
Strata
Making Data Work

Change the world with data. We'll show you how.

strataconf.com

Co-presented by O'REILLY® cloudera®

Strata CONFERENCE + HADOOP WORLD™

Oct 28 – 30, 2013
New York, NY

O'REILLY®

Strata CONFERENCE

Making Data Work

Nov 11 – 13, 2013
London, England

O'REILLY®

Strata CONFERENCE

Making Data Work

Feb 11 – 13, 2014
Santa Clara, CA

O'REILLY®

Strata CONFERENCE Rx

Data Makes a Difference

April 23–25, 2014
Boston, MA

O'REILLY®

Spreading the knowledge of innovators.

The Culture of Big Data

Mike Barlow

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

The Culture of Big Data

by Mike Barlow

Printed in the United States of America.

Published by O’Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O’Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://my.safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editor: Mike Loukides

September 2013: First Edition

Revision History for the First Edition:

2013-10-01: First release

While every precaution has been taken in the preparation of this book, the publisher and authors assume no responsibility for errors or omissions, or for damages resulting from the use of the information contained herein.

ISBN: 978-1-449-36752-7

[LSI]

Table of Contents

The Culture of Big Data Analytics

It's Not Just About Numbers

Today's conversational buzz around big data analytics tends to hover around three general themes: technology, techniques, and the imagined future (either bright or dystopian) of a society in which big data plays a significant role in everyday life.

Typically missing from the buzz are in-depth discussions about the people and processes—the cultural bedrock—required to build viable frameworks and infrastructures supporting big data initiatives in ordinary organizations.

Thoughtful questions must be asked and thoroughly considered. Who is responsible for launching and leading big data initiatives? Is it the CFO, the CMO, the CIO, or someone else? Who determines the success or failure of a big data project? Does big data require corporate governance? What does a big data project team look like? Is it a mixed group of people with overlapping skills or a hand-picked squad of highly trained data scientists? What exactly is a data scientist?

Those types of questions skim the surface of the emerging cultural landscape of big data. They remind us that big data—like other so-called technology revolutions of the recent past—is also a cultural phenomenon and has a social dimension. It's vitally important to remember that most people have not considered the immense difference between a world seen through the lens of a traditional relational database system and a world seen through the lens of a Hadoop Distributed File System.

This paper broadly describes the cultural challenges that invariably accompany efforts to create and sustain big data initiatives in a global economy that is increasingly evolving toward the Hadoop perspective, but whose data-management processes and capabilities are still rooted firmly in the traditional architecture of the data warehouse.

The cultural component of big data is neither trivial nor free. It is not a list of "feel-good" or "fluffy" attributes that are posted on a corporate website. Culture (i.e., people and processes) is integral and critical to the success of any new technology deployment or implementation. That fact has been demonstrated repeatedly over the past six decades of technology evolution. Here is a very brief and incomplete list of recent "technology revolutions" that have radically transformed our social and commercial worlds:

- The shift from vacuum tubes to transistors
- The shift from mainframes to client servers and then to PCs
- The shift from written command lines to clickable icons
- The introduction and rapid adoption of enterprise resource planning (ERP), ecommerce, sales force automation, and customer relationship management (CRM) systems
- The convergence of cloud, mobile, and social networking systems

Each of those revolutions was followed by a period of intense cultural adjustment as individuals and organizations struggled to capitalize on the many benefits created by the newer technologies. It seems unlikely that big data will follow a different trajectory. Technology does not exist in a vacuum. In the same way that a plant needs water and nourishment to grow, technology needs people and process to thrive and succeed.

According to Gartner, 4.4 million big data jobs will be created by 2014, and only a third of them will be filled. Gartner's prediction evokes images of "gold rush" for big data talent, with legions of hardcore quants converting their advanced degrees into lucrative employment deals. That scenario promises high times for data analysts in the short term, but it obscures the longer-term challenges facing organizations that hope to benefit from big data strategies.

Hiring data scientists will be the easy part. The real challenge will be integrating that newly acquired talent into existing organizational structures and inventing new structures that will enable data scientists to generate real value for their organizations.

Playing By the Rules

Misha Ghosh is global solutions leader at MasterCard Advisors, the professional services arm of MasterCard Worldwide. It provides real-time transaction data and proprietary analysis, as well as consulting and marketing services. It's fair to say that MasterCard Advisors is a leader in applied data science. Before joining MasterCard, Ghosh was a senior executive at Bank of America, where he led a variety of data analytics teams and projects. As an experienced practitioner, he knows his way around the obstacles that can slow or undermine big data projects.

"One of the main cultural challenges is securing executive sponsorships," says Ghosh. "You need executive-level partners and champions early on. You also need to make sure that the business folks, the analytic folks, and the technology folks are marching to the same drumbeat."

Instead of trying to stay "under the radar," Ghosh advises big data leaders to play by the rules. "I've seen rogue big data projects pop up, but they tend to fizzle out very quickly," he says. "The old adage that it's better to seek forgiveness afterward than to beg for permission doesn't really hold for big data projects. They are simply too expensive and they require too much collaboration across various parts of the enterprise. So you cannot run them as rogue projects. You need executive buy-in and support."

After making the case to the executive team, you need to keep the spark of enthusiasm alive among all the players involved in supporting or implementing the project. "It's critical to maintain the interest and attention of your constituency. After you've laid out a roadmap of the project so everyone knows where they are going, you need to provide them with regular updates. You need to communicate. If you stumble, you need to let them know why you stumbled and what you will do to overcome the barriers you are facing. Remember, there's no clear path for big data projects. It's like Star Trek—you're going where no one has gone before."

At present, there is not a standard set of best practices for managing big data teams and projects. But an ad hoc set of practices is emerging. "First, you must create transparency. Lay out the objectives. State explicitly what you intend to accomplish and which problems you intend to solve. That's absolutely critical. Your big data teams must be 'use case-centric.' In other words, find a problem first and then solve it.

That seems intuitive, but I've seen many teams do exactly the opposite: first they create a solution and then they look for a problem to solve."

Marcia Tal pioneered the application of advanced data analytics to real-world business problems. She is best known in the analytics industry for creating and building Citigroup's Decision Management function. Its charter was seeking significant industry breakthroughs for growth across Citigroup's retail and wholesale banking businesses. Starting with three people in 2001, Tal grew the function into a scalable organization with more than 1,000 people working in 30 countries. She left Citi in 2011 and formed her own consulting company, Tal Solutions LLC.

"Right now, everyone focuses on the technology of big data," says Tal. "But we need to refocus our attention on the people, the processes, the business partnerships, revenue generation, P&L impact, and business results. Most of the conversation has been about generating insights from big data. Instead we should be talking about how to translate those insights into tangible business results."

Creating a sustainable analytics function within a larger corporate entity requires support from top management, says Tal. But the strength and quality of that support depends on the ability of the analytics function to demonstrate its value to the corporation.

"The organization needs to see a revenue model. It needs to perceive the analytics function as a revenue producer, and not as a cost center. It needs to see the value created by analytics," says Tal. That critical shift in perception occurs as the analytics function forms partnerships with business units across the company and consistently demonstrates the value of its capabilities.

"When we started the Decision Management function at Citi, it was a very small group and we needed to demonstrate our value to the rest of the company. We focused on specific business needs and gaps. We closed the gaps, and we drove revenue and profits. We demonstrated our ability to deliver results. That's how we built our credibility," says Tal.

Targeting specific pain points and helping the business generate more revenue are probably the best strategies for assuring ongoing investment in big data initiatives. "If you aren't focusing on real pain points, you're probably not going to get the commitment you need from the company," says Tal.

No Bucks, No Buck Rogers

Russ Cobb, Vice President of Marketing and Alliances at SAS, also recommends shifting the conversation from technology to people and processes. "The cultural dimension potentially can have a major impact on the success or failure of a big data initiative," says Cobb. "Big data is a hot topic, but technology adoption doesn't equal ROI. A company that doesn't start with at least a general idea of the direction it's heading in and an understanding of how it will define success is not ready for a big data project."

Too much attention is focused on the cost of the investment and too little on the expected return, says Cobb. "Companies try to come up with some measure of ROI, but generally, they put more detail around the 'I' and less detail around the 'R.' It is often easier to calculate costs than it is to understand and articulate the drivers of return."

Cobb sees three major challenges facing organizations with big plans for leveraging big data. The first is not having a clear picture of the destination or desired outcome. The second is hidden costs, mostly in the area of process change. The third and thorniest challenge is organizational. "Are top and middle managers ready to push their decision-making authority out to people on the front lines?" asks Cobb. "One of the reasons for doing big data is that it moves you closer to real-time decision making. But those kinds of decisions tend to be made on the front lines, not in the executive suite. Will management be comfortable with that kind of cultural shift?"

Another way of phrasing the question might be: Is the modern enterprise really ready for big data? Stephen Messer, cofounder and vice chairman of Collective[i], a software-as-a-service business intelligence solution for sales, customer service, and marketing, isn't so sure. "People think this is a technological revolution, but it's really a business revolution enabled by technology," says Messer. Without entrepreneurial leadership from the business, big data is just another technology platform.

"You have to start with the business issue," says Messer. "You need a coalition of people inside the company who share a business problem that can be solved by applying big data. Without that coalition, there is no mission. You have tactics and tools, but you have no strategy. It's not transformational." Michael Gold, CEO of Farsite, a data analytics firm whose clients include Dick's Sporting Goods and the Ohio State

University Medical Center, says it's important to choose projects with manageable scale and clearly defined objectives.

"The questions you answer should be big enough and important enough for people to care," says Gold. "Your projects should create revenue or reduce costs. It's harder to build momentum and maintain enthusiasm for long projects, so keep your projects short. Manage the scope, and make sure you deliver some kind of tangible results."

At a recent Strata + Hadoop World conference in New York, Gold listed three practical steps for broadening support for big data initiatives:

1. Demonstrate ROI for a business use case.
2. Build a team with the skills and ability to execute.
3. Create a detailed plan for operationalizing big data.

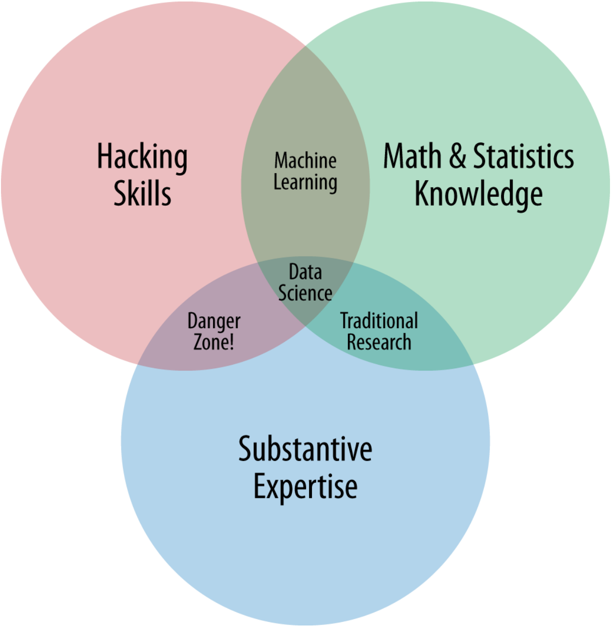
"From our perspective, it's very important that all of the data scientists working on a project understand the client's strategic objectives and what problems we're trying to solve for them," says Gold. "Data scientists look at data differently (and better, we think) when they're thinking about answering a business question, not just trying to build the best analytical models."

It's also important to get feedback from clients early and often. "We work in short bursts (similar to a scrum in an Agile methodology) and then present work to clients so they can react to it," says Gold. "That approach ensures that our data scientists incorporate as much of the clients' knowledge into their work as possible. The short cycles require our teams to be focused and collaborative, which is how we've structured our data science groups."

Operationalizing Predictability

The term "data scientist" has been used loosely for several years, leading to a general sense of confusion over the role and its duties. A headline in the October 2012 edition of the *Harvard Business Review*, "Data Scientist: Sexiest Job of the 21st Century," had the unintended effect of deepening the mystery.

In 2010, Drew Conway, then a Ph.D. candidate in political science at New York University, created a Venn diagram showing the overlapping skill sets of a data scientist. Conway began his career as a computational social scientist in the US intelligence community and has become an expert in applying computational methods to social and behavioral problems at large scale.



From Conway's perspective, a data scientist should possess the following:

1. Hacking skills
2. Math and statistical knowledge
3. Substantive expertise

All three areas are important, but not everyone is convinced that one individual has to embody all the skills of a data scientist to play a useful role on a big data analytics team.

The key to success, as Michael Gold suggested earlier, is operationalizing the processes of big data. Taking it a step further, it is also important to demystify big data. While the *Harvard Business Review* certainly meant no harm, its headline had the effect of glamorizing rather than clarifying the challenges of big data.

Zubin Dowlaty, vice president of innovation and development at Mu Sigma, a provider of decision science services, envisions a future in which big data has become so thoroughly operationalized and automated that humans are no longer required.

"When I walk into an enterprise today, I see the humans are working at 90 percent capacity and the machines are working at 20 percent capacity," says Dowlaty. "Obviously, the machines are capable of handling more work. Machines, unlike humans, scale up very nicely."

Automation is a necessary step in the development of large-scale systems that feed on big data to generate real-time predictive intelligence. "Anticipation denotes intelligence," says Dowlaty, quoting a line from the science-fiction movie *The Fifth Element*. "Operationalizing predictability is what intelligence is all about."

Assembling the Team

At some point in the future, probably sooner rather than later, Dowlaty's vision of automated big data analytics will no doubt become reality. Until then, however, organizations with hopes of leveraging the potential of big data will have to rely on humans to get the work done.

In a 2012 paper,[1] Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer presented the results of interviews with 35 data analysts working in commercial organizations in healthcare, retail, finance, and social networking. Hellerstein, a professor at UC Berkeley, summarized key findings of the paper at a recent Strata Conference. The paper includes insights and models that will likely prove useful to anyone tasked with assembling a big data analytics team.

Based on the interviews, the researchers perceive three basic analyst archetypes:

1. Hacker
2. Scripter
3. Application user

The hacker is typically a fluent programmer and manipulator of data. The scripter performs most of his work within an existing software package and works mostly on data that has been retrieved from a data warehouse by IT staff. The application user relies on spreadsheets or

[1] *Enterprise Data Analysis and Visualization: An Interview Study*.

highly specialized applications and typically works on smaller data sets than hackers and scripters.

It is important for management to understand the differences between those types of analysts when staffing a data analytics team. Hackers are more likely to have a background in computer science. “They are folks who have good facility with programming and systems, but less facility with stats and some of the more ‘scientific’ aspects of data science. They also tend to have less contextual knowledge of the domain-specific questions being explored in the data,” explains Hellerstein.

Scripters, on the other hand, are more likely to be trained statisticians, and app users are more likely to be business people. At the risk of oversimplification, a chart showing the three kinds of analysts and their typical academic backgrounds might look something like this:

Analyst type	Training or academic background
Hacker	Computer science major
Scripter	Statistics major
Application user	MBA

“No (single) one of these categories is more likely than another to succeed on its own,” says Hellerstein. “You can teach stats and business to a hacker, or you can teach computer science and business to a scripter, or you can teach stats and computer science to an app user.”

Scripters and app users would likely require some sort of self-service software to function without help from IT. Similar software might also be useful for hackers, sparing them the drudgery of data prep.

The good news is that several companies are working hard at developing self-service tools that will help analysts become more self-reliant and less dependent on IT. As the tools become more sophisticated and more widely available, it is possible that the distinctions between the three types of analysts might fade or at least become less problematic.

Even when a full suite of practical self-service tools becomes available, it might still make sense to hire a variety of analyst types. For instance, an analytics group that only hired hackers would be like a baseball team that only signed pitchers. Successful teams—whether in business or in sports—tend to include people with various skills, strengths, and viewpoints. Or to put it more bluntly, good luck trying to manage an analytics team made up solely of hackers.

The paper also describes five high-level tasks of data analysis:

1. Discovery
2. Wrangling
3. Profiling
4. Modeling
5. Reporting

Each of the five tasks has a different workflow, presents a different set of challenges or pain points, and involves a different set of tools. Clearly, the universe of practical analytics is a blend of various tasks, tools, and workflows. More to the point, each stage of the analytics process requires an analyst or analysts with particular skills and a particular mindset.

Not all data analysts are created equal, nor are they likely to share the same zeal for different parts of the process. Some analysts will be better at some aspects of analysis than others. Putting together and managing teams that can handle all the necessary phases of data analysis is a major part of the cultural challenge facing organizations as they ramp up big data initiatives.

Team leadership is another challenge. MasterCard's Ghosh recommends that big data projects "be led by passionate and creative data scientists, not by bureaucrats or finance professionals." Others argue that big data initiatives should be led by seasoned corporate executives with boardroom negotiating skills and a keen understanding of how the C-suite operates.

Some companies have hired a chief analytics officer or created an enterprise analytics group that functions as a shared service, similar to an enterprise IT function. Most companies, however, embed analysts within separate business units.

The advantage of planting analysts in individual business units is that it puts the analysts closer to customers and end users. The downside of spreading analytic expertise among various units includes poor communication, lack of collaboration, and the tendency to reinvent the wheel to solve local problems instead of seeking help from other parts of the enterprise.

Another problem with the decentralized analytics model is lack of governance. Today, it is unusual to find the words "governance" and "analytics" in the same sentence. As big data takes on a higher profile in modern corporations, governance will almost certainly become an issue.

For example, very few data analysts save code or models that do not result in practical solutions to immediate problems. As a consequence, analysts can waste an incredible amount of effort making the same or similar mistakes. Unlike, say, chemistry or biology, in which the results of all experiments are duly noted and logged whether or not they are successful, the precise details of data science experiments are usually captured when the analyst succeeds at solving the particular problem at hand.

Another issue that arises from using Hadoop and other frameworks for handling large amounts of unstructured data is the preservation of documentation and potentially important details about the data.

Sean Kandel, a coauthor of the study referenced earlier, sees the "impulse to dump data into an HDFS" as a growing cultural challenge. "When you have to have a traditional data warehousing environment, there is more of a culture around governance and making sure the data that comes in is well structured and fits the global schema," says Kandel. "When you get away from those established practices, it becomes harder to work with the data."

As Kandel and his coauthors write in their paper:

> With relational databases, organizations typically design a database schema and structure incoming data upon load. This process is often time-consuming and difficult, especially with large complex data sets. With Hadoop, analysts typically take advantage of its ability to operate on less structured data formats. Instead of structuring the data up front during ingest, organizations commonly dump data files into the Hadoop Distributed File System (HDFS) with little documentation. Analysis of this data then requires parsing the data during MapReduce jobs or bulk reformatting to load into relational databases. While remaining unstructured, the data may be difficult to search and profile due to the lack of a defined schema. In some cases, the analysts who originally imported and understood the data may no longer work at the company or may have forgotten important details.

"In a large company," says Kandel, "those people might be hard to find. Now you have some interesting questions: Who is responsible for annotating data? How do you structure the data warehouse? How do you convince people to take the time to label the data properly?"

The lack of a disciplined process—what some would call governance —for handling data at every stage of the analytics process suggests the need for automated systems that capture keystrokes or create audit trails that would make it possible for data scientists to replicate or reexamine the work of other data scientists.

Fitting In

Paul Kent is vice president of big data at SAS, one of the earliest and best-known makers of data analytics. He sees a sort of natural "give and take" between traditional analysts working with limited sets of structured data and a newer generation of analysts who seem comfortable handling an endless deluge of unstructured data.

"I think you have to give the newer analysts their own space. They'll need to preserve some of their independence. They won't be happy playing by the old school rules," says Kent. "Big data has changed the way we look at data. It's messy, and it's not expensive to save. So we save as much as we can. And when we have questions in the future, we'll map those questions to the data that we've saved."

In the past, data infrastructures were designed around a known set of questions. Today, it's much harder to predict the questions that will be asked. That uncertainty makes it nearly impossible to build traditional-style infrastructures for handling big data.

"We really can't design the perfect structure for data and then just pour data into it," says Kent. "So you have to think about it the other way around. We don't even know the questions we're going to ask tomorrow or next month. So we keep as much data as we can and we try to be as flexible as possible so we can answer questions when they come up."

The "old school" perspective was that "if you think real hard, you can design a nice structure for your data and then fill it up whenever you get your data—every week, every day, or every hour," says Kent. If the structure you designed was good enough, it could be tweaked or modified over time to keep up with the changing needs of the market.

"The new school says, 'Nope, that won't work. Let's just save the data as it comes in. We'll merge it and join it and splice it on a case-by-case basis.' The new school approach doesn't necessarily need a relational database. Sometimes they'll just work with raw files from the originating system," says Kent. Andreas Weigend teaches at Stanford University and directs the Social Data Lab. The former chief scientist at Amazon, he helped the company build the customer-centric, measurement-focused culture that has become central to its success. Weigend sees data-driven companies following an evolutionary path from "data set to tool set to skill set to mindset." He suggests eight basic rules for organizations in search of a big data strategy:

1. Start with the problem, not with the data.
2. Share data to get data.
3. Align interests of all parties.
4. Make it trivially easy for people to contribute, connect, collaborate.
5. Base the equation of your business on customer-centric metrics.
6. Decompose the business into its "atoms."
7. Let people do what people are good at, and computers what computers are good at.
8. Thou shalt not blame technology for barriers of institutions and society.

Weigend's list of rules focuses entirely on the cultural side of big data. In some ways, it's like the driver's manual you read in high school: heavy on driving etiquette and light on auto mechanics. The miracle of the internal combustion engine is taken for granted. What matters now is traveling safely from Point A to Point B.

Conversations about big data have moved up the food chain. People seem less interested in the technical details and more interested in how big data can help their companies become more effective, more nimble, and more competitive. As Marcia Tal puts it, "The C-suite wants to know what big data is worth to the organization. They want to see the revenue it generates. They want to understand its value and measure the return on their investment."

About the Author

Mike Barlow is an award-winning journalist, author, and communications strategy consultant. Since launching his own firm, Cumulus Partners, he has represented major organizations in numerous industries. Mike is coauthor of *The Executive's Guide to Enterprise Social Media Strategy* (Wiley, 2011) and *Partnering with the CIO: The Future of IT Sales Seen Through the Eyes of Key Decision Makers* (Wiley, 2007). He is also the writer of many articles, reports, and white papers on marketing strategy, marketing automation, customer intelligence, business performance management, collaborative social networking, cloud computing, and big data analytics. Over the course of a long career, Mike was a reporter and editor at several respected suburban daily newspapers, including *The Journal News* and the *Stamford Advocate*. His feature stories and columns appeared regularly in *The Los Angeles Times*, *Chicago Tribune*, *Miami Herald*, *Newsday*, and other major US dailies.