Cutter IT Journal

The Journal of Information Technology Management

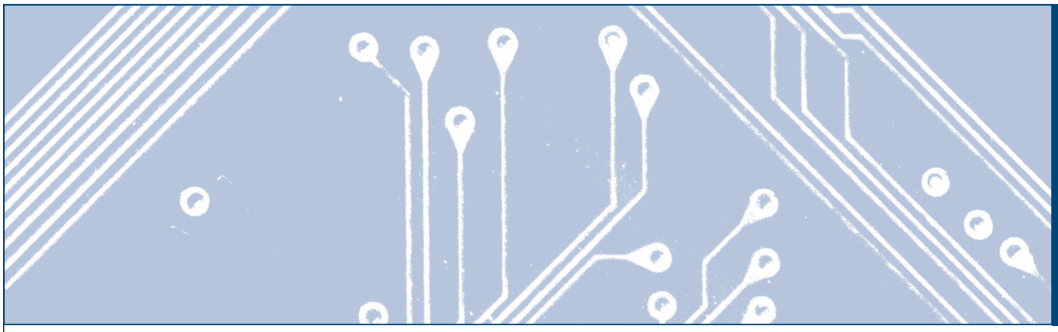
Vol. 29, No. 6, June 2016

"Of all the adjectives attached to *analytics* in the market, it is *predictive* that gains most attention. This is because the actual business value of collecting and analyzing data is closely related to — and arguably depends on — the ability of business to forecast future events, outcomes, or behaviors."

— Barry Devlin, Guest Editor

Cultivating Success in Big Data Analytics

CUTTER CONSORTIUM



Cutter IT Journal

About Cutter IT Journal

Part of Cutter Consortium's mission is to foster debate and dialogue on the business technology issues challenging enterprises today, helping organizations leverage IT for competitive advantage and business success. Cutter's philosophy is that most of the issues that managers face are complex enough to merit examination that goes beyond simple pronouncements. Founded in 1987 as *American Programmer* by Ed Yourdon, *Cutter IT Journal* is one of Cutter's key venues for debate.

The monthly *Cutter IT Journal* and its companion *Cutter IT Advisor* offer a variety of perspectives on the issues you're dealing with today. Armed with opinion, data, and advice, you'll be able to make the best decisions, employ the best practices, and choose the right strategies for your organization.

Unlike academic journals, *Cutter IT Journal* doesn't water down or delay its coverage of timely issues with lengthy peer reviews. Each month, our expert Guest Editor delivers articles by internationally known IT practitioners that include case studies, research findings, and experience-based opinion on the IT topics enterprises face today — not issues you were dealing with six months ago, or those that are so esoteric you might not ever need to learn from others' experiences. No other journal brings together so many cutting-edge thinkers or lets them speak so bluntly.

Cutter IT Journal subscribers consider the *Journal* a "consultancy in print" and liken each month's issue to the impassioned debates they participate in at the end of a day at a conference.

Every facet of IT — application integration, security, portfolio management, and testing, to name a few — plays a role in the success or failure of your organization's IT efforts. Only *Cutter IT Journal* and *Cutter IT Advisor* deliver a comprehensive treatment of these critical issues and help you make informed decisions about the strategies that can improve IT's performance.

Cutter IT Journal is unique in that it is written by IT professionals — people like you who face the same challenges and are under the same pressures to get the job done. *Cutter IT Journal* brings you frank, honest accounts of what works, what doesn't, and why.

Put your IT concerns in a business context. Discover the best ways to pitch new ideas to executive management. Ensure the success of your IT organization in an economy that encourages outsourcing and intense international competition. Avoid the common pitfalls and work smarter while under tighter constraints. You'll learn how to do all this and more when you subscribe to *Cutter IT Journal*.

Cutter IT Journal®

Cutter Business Technology Council:
Rob Austin, Ron Blitstein, Tom DeMarco, Lynne Ellyn, Vince Kellen, Tim Lister, Lou Mazzucchelli, and Robert D. Scott

Founding Editor: Ed Yourdon
Publisher: Karen Fine Coburn
Group Publisher: Chris Generali
Managing Editor: Karen Pasley
Production Editor: Linda Dias
Client Services: service@cutter.com

Cutter IT Journal® is published 12 times a year by Cutter Information LLC, 37 Broadway, Suite 1, Arlington, MA 02474-5552, USA (Tel: +1 781 648 8700; Fax: +1 781 648 8707; Email: citjeditorial@cutter.com; Website: www.cutter.com; Twitter: @cuttertweets; Facebook: Cutter Consortium). Print ISSN: 1522-7383; online/electronic ISSN: 1554-5946.

Subscription rates are US $485 a year in North America, US $585 elsewhere, payable to Cutter Information LLC. Reprints, bulk purchases, past issues, and multiple subscription and site license rates are available on request.

Opening Statement

by Barry Devlin, Guest Editor

"Big data" and "analytics" are among the most overhyped and abused terms in today's IT lexicon. Despite widespread use for almost a decade, their precise meanings remain mysterious and fluid. It is beyond doubt that the volume of data being generated and gathered has been growing exponentially and will continue to do so, intuitively validating the *big* moniker. However, other vital characteristics of today's data, such as structure, transience, and — most disturbingly — meaning and value, remain highly ambiguous. Analytics also remains troublingly vague, as it is prefixed with adjectives ranging from *operational* to *predictive*.

In such circumstances, defining what success in big data analytics (BDA) might mean is problematic. Describing how it could be cultivated would seem especially challenging. Nonetheless, that is what this issue of *Cutter IT Journal* seeks to do, beginning with the premise that, irrespective of definitional difficulties, success in implementing BDA is predicated on addressing four specific aspects of the overall process:

- **People.** Data scientists have been called unicorns due to their elusive nature. Finding them externally or growing them internally and defining mandatory skills, roles, and responsibilities are among the challenges organizations face.
- **Preparation.** Sourcing, cleansing, and contextualizing incoming data before analysis can make the difference between valid discoveries and rabid nonsense. With such preparation reportedly taking 80% of the time and effort of data scientists, new approaches are needed to streamline the process.
- **Prediction.** As the industry progresses from business intelligence and hand-crafted analysis to self-improving deep learning algorithms, issues of understanding, control, and trustworthiness will need to be addressed.
- **Production.** Procedures for effective and seamless transition from discoveries in the analytic laboratory to action in the production environments of manufacturing, operations, and sales will be vital.

These four aspects neither stand alone as all that you must address, nor are they unique to big data analytics. You will need to take care of all the other, usual pieces of process and project management involved in any significant IT undertaking. However, one or (usually) more of these four aspects present particular difficulties in most BDA implementations.

Teaming Up the Right People

BDA has a long history in the open source software scene, particularly focused around Apache Hadoop and its accompanying menagerie, and an association with the Web behemoths of big data, such as Google, Facebook, and Amazon. It was within this environment that the job title of "data scientist" was popularized and from here that it has become the allegedly sexiest job in the world. Many people now want to be one, and many companies want one or more of them. Recruiters declare huge shortages, and authors have created lists of diverse characteristics for the role, ranging from degrees in statistics to Perl programming skills, and from experience in presenting to and influencing management to deep knowledge of business data and processing.

Such guidance is undoubtedly valuable but may be offset by advice from vendors and consultants to avoid "polluting" your future-oriented data scientist team with "BInosaurs" — staff who have grown up in the traditional business intelligence (BI) environment. This is unfortunate. Many companies that are transitioning from a purely physical business environment to the digitized world face the challenge of crafting a functioning team that can extract the undisputed value that exists in big data while also creating and maintaining a data management environment where data quality and governance are critical. Engaging the right people is important, but creating an integrated and empowered team is the first and most vital step toward success in BDA.

Preparing the Ground

In data warehousing, this step used to go under the label of ETL — extract, transform, and load — and it was renowned as the most challenging aspect of

building a BI system. Data in existing operational sources was often not how it was described in the specs. Even the specs were often not what they seemed. Data that should match across sources didn't. The list goes on, but the result was that preparing data for consumption in the BI system was the part of the project that consumed the most resources and could be almost guaranteed to overrun.

Fast-forward to big data analytics, and it appears that history is doomed to repeat itself. Despite the fact that the majority of big data is sourced externally to the enterprise, coming from notoriously uncertified social media sources and highly unreliable Internet of Things (IoT) sensors, data scientists express ongoing incredulity that data preparation (now sexily called "wrangling") takes such time and effort. According to Monica Rogati, VP for data science at Jawbone, "Data wrangling is a huge — and surprisingly so — part of the job. At times, it feels like everything we do."[1]

Since then, a number of vendors have introduced or expanded offerings that address preparation, cleansing, wrangling, and quality of big data in the Hadoop environment. This is, of course, welcome. However, it remains a technical, product-level solution to a broader problem.

Successful BDA requires a fundamental rethinking of the process of data preparation. This begins with a policy decision that only data of known and agreed levels of quality can be introduced into particular analyses. Some initial analyses may use "dirtier" data, of course, but as the process progresses toward actual decision making, more stringent requirements on meaning, structure, and completeness may be mandated. Modeling of data, both in advance of ingestion and on an ongoing, in-stream basis, must become the norm. Rules that limit mixing or combining data of different levels of quality will be required and must be enforced. For example, at the most obvious level, combining data from social media sources with regulated financial data would be disallowed. "Health warnings" should be attached to input and output data sets, as well as analytical reports, clearly stating the business process or circumstances under which these sets may or may not be used.

Predicting the Future Is Hard

Of all the adjectives attached to *analytics* in the market, it is *predictive* that gains most attention. This is because the actual business value of collecting and analyzing data is closely related to — and arguably depends on — the ability of business to forecast future events, outcomes, or behaviors. As a result, the term "analytics" has largely displaced "business intelligence" in the market over the past decade.

In common usage, the meaning of [*choose an adjective*] analytics ranges from basic data query and reporting, through statistical analysis, to the application of advanced AI techniques to decision-making support. One useful approach to clearing some of the confusion is to look at the purpose and time frame of the analytic activity. This leads to five classes of analytics:

1. **Descriptive** — focuses on the past to describe what happened at a detailed and/or summary level, corresponding to traditional BI (query and reporting) and data mining (statistics)
2. **Operational** — focuses on the present moment, often down to subsecond intervals, and seeks to know what is currently happening in great detail in real time
3. **Diagnostic** — spans the past and present time frames to understand why the things discovered in descriptive and operational analytics actually occurred, to deduce causation
4. **Predictive** — focuses on the future in an effort to forecast what may happen with some level of statistical probability
5. **Prescriptive** — takes input from the previous four types of analytics and attempts to influence future behaviors and events, using optimization and simulation techniques, for example

Different authors use different subsets of the above list and use their lists for different purposes. For example, these categories can be used to evaluate tool and product capabilities when comparing vendor solutions. Alternatively, one can describe an organization's maturity in decision-making support by observing their capabilities in these types of analytics, understanding that the classes build one upon the other in order listed

above (i.e., descriptive is the most basic and prescriptive the most advanced).

It is useful to note the overlap and interdependence of the categories listed above. Success in predicting behavior or outcomes is built upon a strong foundation of descriptive work, both BI and data mining, as well as — in many cases — serious operational, real-time analytics. Beneath both these aspects lies a firm foundation of data management and quality work, of which building and maintaining an enterprise data warehouse (EDW) infrastructure is most important.

The role of the EDW in prediction and, indeed, prescription, is to create and manage *core business information*, the legally binding record of the state and history of the business. The task of forecasting the future can be eased only if this core information is of high quality. Only then can the business have confidence that the results obtained have a high probability of being valid.

The challenge now emerging — and doing so rapidly — is to understand and address the implications of models and algorithms that are capable of self-improvement. Rapid advances in a range of overlapping fields such as deep learning, AI, and cognitive computing are being incorporated in predictive and prescriptive analytics. Whether implemented through automation (replacement of human decision makers), augmentation (collaboration between humans and machines according to their respective strengths), or, most likely, a combination of both, new processes and models are urgently required, and new legislative and ethical frameworks will need to be devised.

Production Is the End Point

The popularity of data scientists and the creation of "analytics labs" in larger organizations have led to a popular image of white-coated researchers pursuing lofty searches for truth in the data lakes of the business world. Unfortunately, this image misleads. Data science falls more correctly under the category of applied R&D rather than that of pure, fundamental research. As with all applied R&D, the aim of data science is not to discover new truths for their own sake, but rather to take discoveries from the lab into day-to-day business situations to improve the bottom line. This transition from exploration to production is usually messy. However, doing it well — in terms of speed, quality, and ease — is the final guarantor of success in BDA.

In traditional, physical industries, R&D and production are very distinct and separate activities, carried out in different places, using different materials and tools, and so on. There exists between them a well-defined boundary with well-formulated procedures and rules for moving from one side to the other. In the case of analytics, this boundary is far from obvious, for both historical and practical reasons.

Historically, IT has created virtually all of the data used by the business; such process-oriented data is essentially a byproduct of automating business processes via computers. In this situation, R&D on such data is largely meaningless. The focus of IT has thus first been on creating and managing the data itself (via operational systems) and, second, making it available for management reporting and problem solving (via informational or BI systems). In both cases, quality and consistency are key. In essence, IT has had no historical reason to differentiate between R&D and production.

The current drive toward digitization of business changes the situation dramatically, with an influx of external data swamping traditional process-oriented data in both physical volume and business attention. Such data, from social media and the IoT, differs significantly from process-mediated data in terms of structure, quality, lifespan, and more. It provides ample opportunities for R&D (analytics), but this is best performed in conjunction with the core business information contained in process-mediated data. This need to blend the two types of data in analytics blurs the boundary between R&D and production in practice.

Formalizing and strengthening this boundary is vital for success in BDA. While technology does have a role to play (through metadata management and data quality tools, for example), this is far from the simplistic formula of "relational databases for production and Hadoop for analytics." New conceptual and logical architectural frameworks that address both modern real-time business needs and today's disparate data types are necessary.[2] In addition, formal methodologies and real-life implementations are emerging that show how this can be achieved.

In This Issue

This issue of *Cutter IT Journal* offers a variety of perspectives on what is required to ensure success in big data analytics, covering topics from methodologies to architectures, as well as a dive into one of the key technologies of the field. Our authors provide five distinctly different views on where you should direct your attention over the coming years of evolving BDA practice.

We open with a thought-provoking article by Steve Bell and Karen Whitley Bell, who use the "lessons learned in

over five decades of Lean Thinking" to consider how to get the most value from BDA. Starting from a consideration of the adaptive learning organization, they take us to a Six P Model that relates purpose, process, and people to planning, performance, and problems, concluding that "paradoxically, to achieve desired outcomes, managers must pay more attention to the process and less to those outcomes."

In our second article, Cutter Senior Consultant Bhuvan Unhelkar combines academic and hands-on experience to offer the Big Data Framework for Agile Business as an architectural foundation for BDA. Arguing that technology must be balanced with a deep appreciation of business drivers and realities, Unhelkar's framework includes "agile values for business, organizational roles in big data, building blocks of big data strategies for business (including the role of analytics within those strategies), key artifacts in big data adoption, business conditions and limitations, agile practices, and a compendium (repository)" as the basis for successful implementation.

Our third offering, from Jeff Carr, dives into a fascinating exploration of the importance of semistructured data and NoSQL technology in support of BDA. He defines a generic data model for NoSQL and describes eight fundamental capabilities a NoSQL analytics system must have to derive analytic value from arbitrary semistructured data. These attributes can form the basis for evaluating the ability of any tool or system to perform generalized NoSQL analytics.

Next, Donald Wynn and Renée Pratt take us back to the organizational challenges to maximizing value in BDA from the viewpoint of innovation. These challenges revolve around managing the implementation of processes, data management, and staffing. They note the resemblance of driving BDA innovation to traditional business process management, describing an iterative process that begins where the previous evaluation phase concluded, which "naturally leads to the contemplation of desired future changes." Of particular interest is their analysis of when top-down and bottom-up approaches to BDA implementation are most appropriate in organizations.

Our fifth and final article, from Mohan Babu K, examines the implementation of big data analytics in an industry seldom associated with IT: agriculture. While most industries have arrived at BDA from a history of business intelligence, agriculture offers a greenfield (pun intended) scenario as IoT sensors provide a completely new foundation for augmenting human decision making among people for whom "analysis of data is not their core competence." Babu K describes a framework for analytics in agriculture that will be familiar to practitioners across all industries, once again demonstrating that analytics applies everywhere.

Despite the IT industry's many years of talking about and implementing big data analytics, the articles in this issue of *Cutter IT Journal* serve to emphasize that this field is still undergoing significant evolution and that there remain widely varying ways of planning and implementing BDA. In some sense, we have only scratched the surface of the field, although we have dug deep in particular areas. I trust you will enjoy reading the articles here and believe you will find some nuggets of inspiration that will help you drive success in your own projects.

Endnotes

[1]Lohr, Steve. "For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights." *The New York Times*, 17 August 2014 (www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html?_r=1).

[2]Devlin, Barry. *Business unIntelligence: Insight and Innovation Beyond Analytics and Big Data*. Technics Publications, 2013.

Barry Devlin is a Senior Consultant with Cutter Consortium's Data Analytics & Digital Technologies practice. Dr. Devlin is a founder of the data warehousing industry, having published the first architectural paper on the topic in IBM Systems Journal *in 1988. He is among the foremost authorities on business intelligence (BI) and data warehousing. With today's emphasis on big data, the Internet of Things (IoT), and the blending of the operational and informational worlds, Dr. Devlin is creating new methods and approaches to extend BI into true business insight.*

As an IT and management consultant, Dr. Devlin has guided BI and big data strategies and implementations in a wide range of business areas over the past 20 years. Clients include major banks, insurance companies, and retailers throughout Europe, the US, and South Africa. Dr. Devlin is a widely respected analyst, consultant, lecturer, and author of the seminal book Data Warehouse: From Architecture to Implementation. *His most recent book is* Business unIntelligence: Insight and Innovation Beyond Analytics and Big Data. *In addition, he is a regular contributor to the industry through blogs, journal articles, and Twitter.*

Dr. Devlin founded 9sight Consulting in 2008. Previously, he was an international consultant, manager, systems architect, software evangelist, and Distinguished Engineer with IBM for more than 20 years. He can be reached at bdevlin@cutter.com.

Big Data and Lean Thinking: Balancing Purpose, Process, and People

by Steve Bell and Karen Whitley Bell

> Data is important, but I prefer facts.
>
> — Taiichi Ohno, originator of the Toyota Production System

In the closing chapter of *The Innovators*, the story of the emergence of the computer age, Walter Isaacson explores the history of computing and speculates on its future. He explains that cognitive computing — once called "artificial intelligence," the notion that computers will be able to "think" — has perpetually "remained a mirage, always about 20 years away."[1]

Suddenly we are witnessing the convergence of many advances — from self-driving vehicles to new medical diagnostic and delivery methods — and the fruition of these breakthroughs may deeply impact us all in unforeseeable ways. While there are many differing views, and even fears, about big data and cognitive computing, there is also tremendous opportunity. As Isaacson concludes, "New platforms, services, and social networks are increasingly enabling fresh opportunities for individual imagination and collaborative creativity. This innovation will come from people who are able to link beauty to engineering ... humanity to technology."[2]

We must make wise choices about how we invest in and use these emerging technologies. So the premise of this article is this: how do we ensure that we are getting the most from big data, cognitive computing, and whatever lies beyond, to improve the probability of making the right decisions, in the right context, and for the right reasons? We believe that lessons learned in over five decades of Lean Thinking can help guide us forward in this journey, and we will use examples from the financial services industry to illustrate them.

The Essence of an Adaptive Learning Organization

From the beginning, there were two branches of artificial intelligence: one that seeks to replace human cognition and one that seeks to augment and complement it, to make it more effective. Of these two paths, Isaacson observes, human-computer symbiosis has been more successful. "The goal is not to replicate human brains," says John Kelly, the director of IBM Research. "Rather, in the era of cognitive systems, humans and machines will collaborate to produce better results, each bringing their own superior skills to the partnership."[3]

Whether computers will ever really "think" may forever remain a philosophical and theological debate. Nevertheless, we can expect that computing will increasingly augment and even replace humans in many circumstances, just as robotics and other technology advances have done in recent memory. So what is the right approach to optimize human-computer interaction?

In 1983, just as the notion of a personal computer was becoming popular, Eiji Toyoda, the Toyota chairman who supported Taiichi Ohno in the creation of the Toyota Production System (the origin of Lean), made this observation:

> Society has reached the point where one can push a button and be immediately deluged with technical and managerial information. This is all very convenient, of course, but if one is not careful, there is a danger of losing the ability to think. We must remember that in the end it is the individual human being who must solve the problems.[4]

A Lean organization encourages every individual to actively seek out problems (rather than avoid or deny them), because problems are the catalyst for continuous improvement and innovation. Lean practice is founded on the scientific method of problem solving; information enables the perpetual feedback loop of continuous improvement and innovation toward the creation of value for the customer. While Lean practice emphasizes data-driven decision making, it must be done with the proper understanding and context, hence the importance of *gemba* — the idea that one must go to the source, to where the work is done, in order to fully understand the situation.

One interpretation of gemba is context and situational awareness through firsthand observation. A Lean management system encourages those closest to the customer, those who best understand the work, to own

their processes, solve their problems, and make their own improvements and innovations, guided by the strategic priorities of the overall organization.

To achieve this new way of thinking and acting, managers must step away from their PowerPoints and spreadsheets, leave their offices and meeting rooms, and actively observe the situation in person. They must overcome the tendency to think they know the answers and learn to rely on the understanding of those who deal with the problems on a daily basis. Their role changes from command and control to coaching, providing support and encouragement to help others overcome obstacles and develop insights into cause and effect, so they can continuously and sustainably improve their work.

Big data analytics can play a vital role in directly engaging individuals and teams at all levels within the enterprise with the virtual voice of the customer, creating an instantaneous line-of-sight view.

Not only does this result in better processes and better problem-solving skills, it nurtures future leaders. Lean requires a whole new style of leadership and management. As Jim Womack (coauthor of *Lean Thinking*) insists, "The Lean manager realizes that no manager at a higher level can or should solve a problem at a lower level; problems can only be solved where they live, by those living with them."[5]

Being closer to the problem is not just a matter of cutting through vertical hierarchy; it requires a contextual shift to the horizontal flow of value. *Value streams* begin with a need expressed by a customer and conclude with delivery and satisfaction, which creates a complete end-to-end perspective. A Lean approach to problem solving must thus begin with a deep understanding of the organizational architecture — breaking down functional silos and their counterproductive attitudes, measures, and incentives in order to create what is often called a "line-of-sight view" to the customer needs and experience.

Consider the nature of value streams in today's marketplace. Teams are often geographically distributed, and customers may be globally dispersed and very diverse in their characteristics and consumption patterns. While "going to gemba" to talk and learn directly with customers will always be valuable in gaining new insights, these face-to-face conversations will be with a very small, often non-random sample of a larger, diverse population and will not provide the breadth of insight that a larger sampling can provide.

Big data analytics can play a vital role in directly engaging individuals and teams at all levels within the enterprise with the *virtual voice of the customer*, creating an instantaneous line-of-sight view. This continuous feedback and learning provides an efficient and rapid way to design and conduct experiments of any scale to explore customer preference and actions, enabling individuals and teams to make decisions that continuously align with and inform enterprise strategy.

Seeing the Whole

People often say that "politics" or "culture" gets in the way of effective decision making. What this usually indicates, in Lean terms, is that individual functions attempt to locally optimize the segment of the value stream for which they have responsibility and by which they are measured and rewarded. This behavior leads to suboptimization, where segments of the value stream may compete with each other, often pushing the problems and waste into another function's area, like pieces on a game board, while not adding value (speed, quality, cost, safety, experience) to the end customer.

While each function typically has abundant data at the start of an improvement effort, this data is usually very messy — it is strongly influenced and filtered by how each value stream segment is measured and rewarded, so collectively the data does not represent the information essential to improve the flow of the overall process. Once data, evidence, artifacts, and anecdotes across the entire end-to-end value stream are gathered in one place, and as the team fully engages in value stream mapping and analysis, together they come to realize that data from each segment reflects a gathering of suboptimal and often counterproductive points of view. At this moment, someone on the team is often heard to exclaim, "I have been working here X years, and this is the first time I truly understand why this process just doesn't work!"

Simply normalizing the composite data does not correct the distorted perspective on the overall flow. The team must together map and analyze the value stream, determining what are the key drivers for improving the end-to-end flow that transcend individual interests, metrics, and incentives. The inherent unreliability of technically normalized data (which creates the appearance of coherence) poses a significant challenge for effective data-driven decision making. Lean thinkers have learned that when a process is not well understood

from end to end, they should not begin with a data analysis deep-dive. First everyone must visualize the overall flow and context of the process, transforming a disparate gathering of stakeholders into a purpose-driven team. With that understanding, the team can together identify and collect the relevant data and prioritize which problems must be solved to optimize that collective, overarching purpose. Then they can design experiments accordingly to test the impact of changes on the overall outcomes.

The team must look at both performance to the overarching goal and the contribution of each value stream segment toward that goal to optimize value and performance for the customer. If this collective "clarity of purpose"[6] isn't deliberately nurtured, suboptimized improvement efforts — while creating the appearance of progress — can actually drive the value stream and its participants deeper into dysfunction and despair. Misguided big data efforts can do that as well, despite best intentions of everyone involved.

What Makes the Whole?

What might be streamlined and automated today may be entirely replaced tomorrow by innovative, disruptive human creativity. As the strategist Arie de Geus observes, "The ability to learn faster than your competitors may be the only sustainable competitive advantage."[7] This is the essence of Lean Thinking and how it can transform the behavior and culture of an enterprise, preparing it to compete in a highly disruptive future.

Machine learning can help people gain new insights and make informed decisions, but people can't always turn the running of a process entirely over to machines, especially within a complex and dynamic system. While a machine may surface hidden signatures that no human would think to look for, we strongly believe that humans must be able to see that result and, at a minimum, hypothesize the mechanics that allow for a relationship between those new Xs and the Ys. This requires special insight and creativity.

If the analytics engine is a black box to the people responsible for the work, imposed on them by experts, it can become a blunt instrument of management command and control. Automated monitoring and sophisticated analytics can help the team to monitor the state of a process, and instantly sense and respond to deviations, but the team must understand their process to understand the meaning of the deviations so they can solve the right problems and make the right decisions to achieve the desired outcomes.

Big data has already demonstrated many successes, and experts assert that cognitive computing systems can actually make the context behind decision making "computable," acting as a proxy for human intuition. It is that convergence — human creativity supported by relevant information — that offers the greatest potential.

From a Lean perspective, then, how and when can we make context computable in order to help those who are closest to the problem? To address this question, let's briefly examine the fundamental problem-solving journey embodied in Lean Thinking. In his article "Purpose, Process, People," Jim Womack observes:

> ... business purpose always has these two aspects — what you need to do better to satisfy your customers and what you need to do better to survive and prosper as a business. Then, with a simple statement of business purpose in hand, it's time to assess the process that provides the value the customer is seeking. Brilliant processes addressing business purpose don't just happen. They are created by teams led by some responsible person. And they are operated on a continuing basis by larger teams led by value stream managers.[8]

If the analytics engine is a black box to the people responsible for the work, imposed on them by experts, it can become a blunt instrument of management command and control.

The further elaboration of Womack's Purpose, Process, and People in Figure 1[9] illustrates the inner feedback looping and learning of continuous improvement, where the people responsible for the process are able to guide its improvement. When an organization is able to incorporate this mindset into their management systems, behavior, and culture, we can say they have become an adaptive learning organization.

A common challenge with traditional management systems, however, is that they attempt to control a complex, dynamic process from the top down and outside in, focusing primarily on outcomes. Trying to reward and punish those doing the work so as to achieve the desired outcomes can lead to tampering and gaming, where individuals and teams hit their quantitative targets but cause negative consequences in other parts of the organization. W. Edwards Deming argued for constancy of purpose and the abandonment of quantitative quotas, numerical goals, and the popular Management by Objectives approach. In their place, he emphasized skilled leadership.[10]

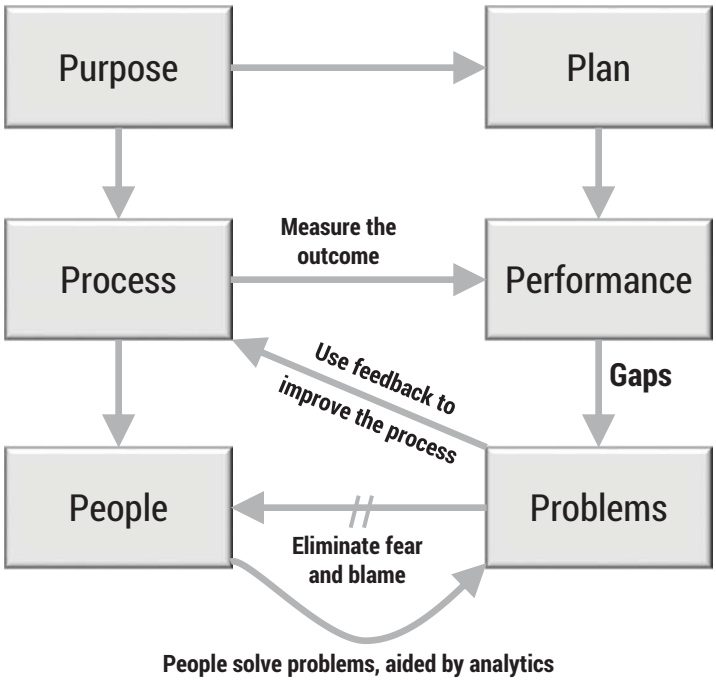


Figure 1 — The Six P Model.

Deeper understanding of the in-process measures can help teams understand the true drivers of performance, regardless of where the responsibilities and accountabilities reside, so they can continuously fine-tune to produce the desired outcomes, especially in a complex and dynamic system. Paradoxically, to achieve desired outcomes, managers must pay more attention to the process and less to those outcomes — it is the internal process drivers that we must analyze, with a keen eye for cause and effect.

Case Studies: Problem First or Data First?

Let's now apply our Lean Thinking insights to better understand two distinct approaches to using big data: problem first and data first, also called *supervised* and *unsupervised* machine learning. The first requires a hypothesis in order to utilize the data, while the second generates hypotheses from raw data. They are both useful for continuous improvement and innovation, but they require a different mindset from the practitioners.

Problem First

The problem first approach requires that we choose a problem, establish why it is important (purpose), then develop hypotheses, which are tested and result in learning (from either a positive or negative test) with the entire hypothesis-testing-learning cycle repeating iteratively as long as resources, demand, and prioritization against other challenges allow.

Hypothesis testing itself requires well-selected measurement points and good, stable data, which may be lacking. The underlying process problem being investigated is usually caused by variation, which contributes to challenges in effective data collection and analysis, especially when voluminous data is gathered through the machine learning process. The team must often take an initial step that focuses on measurement and data collection proofs-of-concept. Often a data cleansing effort is also needed to ensure the signal is clear (noise is minimized). The data itself, and the team that owns the process that creates/utilizes it, must essentially go through a mini-lifecycle to establish a baseline before the data can be used to guide the team toward continuous improvement.[11]

In the financial services industry, there are many specific challenges regarding timely deliveries of daily data needed to enable customer market activities. One very large financial services firm embarked on an effort to improve the delivery of a daily market calculation through enhanced automation and increased Straight Through Processing (STP), measured by how many (or few) times the calculation train stopped for human intervention. There was a strong perception within the team, and by management, that most late deliveries were caused by technology incidents. While the technologies were complex and challenging to integrate,

the only measurements taken that could identify bottlenecks were post-mortem debriefs using subjective assessments — no real data was captured for evaluation of the process.

After just a few months of measurement, visualization, evaluation, modeling, and reporting, things started becoming clear to the process owner. First, as the model revealed the most sensitive process steps, it was apparent that technology was not the primary cause of missed deliveries, as some had originally asserted. Second, there was a specific technology bottleneck that merited analysis. Due to the vast complexity of this single step, the team applied supervised machine learning, and the results were surprising.

When a single, unexpected transaction was revealed to create an excessive processing bottleneck, the team investigated the predecessor steps and rationale for that transaction. They discovered that there was a misunderstanding among the operations staff about how to best gauge technology process status — was the process complete or not? Frequent inquiry and process restarts were causing vastly increased system utilization, unnecessarily slowing down everyone's calculations. This discovery helped the team develop a simple, fast, and noninvasive way to determine process status, which the users quickly adopted, increasing the available time between task completion and due date by 40%. This change led to improved delivery reliability for the customers and is expected to reduce operating expenses due to reduced errors, failure demand, and the resulting rework.

In this case, the team identified a problem, and through rigorous understanding of the process and collection and examination of the data that supported the process, they realized that their unquestioned assumptions were incorrect. This helped them to improve the process and the outcomes.

Data First

The data first approach is very different in nature. Computers are ideal for spotting hidden patterns and relationships that humans can't see. But comprehending the relationships, in order to produce desired outcomes, requires context and situational understanding.

Another financial services case demonstrates the value of using computing power to reduce measurement intervals to provide actionable data. One value stream being measured was already achieving a high STP percentage, but the delivery against the customer satisfaction specification was still lagging. The technology team had a strong *sense* of where the problems were, but the team members were reacting to stale data that lacked client specificity in an area where much of the technology in each value stream is client-specific. While there were daily, even hourly, incidents that the team was responding to in a reactive mode, the overall performance of the value stream was only evaluated on a monthly basis due to the challenges associated with collecting, managing, and reporting on aggregated data for a continuous flow process servicing so many clients.

Computers are ideal for spotting hidden patterns and relationships that humans can't see. But comprehending the relationships, in order to produce desired outcomes, requires context and situational understanding.

Here the challenge was to first solve the data and measurement problem. The team introduced standardization of measurements and automation of the data collection and reporting activities, reducing the monthly reporting lag from 20 days to five, and creating a daily report for the technology team, who consumed this information during problem-solving portions of their daily stand-up meetings. In parallel, a model was developed using data science to understand the process steps that were consuming the most time.

With their new ability to see emergent patterns, application developers were able to spot — and in some cases even anticipate — important events. By focusing on certain client sets and their specific technology architecture, they were able to quickly respond to deviations, thus establishing a clear connection between cause and effect, which helped them to drive out unnecessary variation. The measurement team also asked the technology team for more specificity in the incident ticket details to allow for better correlation between the user experience of process performance and the modeled behavior. This, along with technology process changes made by the subject matter experts, for example, reduced the delivery failure rate by 65% in one key client area alone.

An additional example of data first learning by this same team is a recent proof of concept using unsupervised machine learning to predict the extent of the client impact based on attributes known at the start of an incident. This is enabling real-time decision making by technology support staff as to where to focus their

attention on rapid resolution of the more significant problems, and it may ultimately lead to preventative, and perhaps even predictive, countermeasures.

Purpose and People First

As we have shown with these business process improvement examples, both problem first and data first approaches can help teams sense and respond to emergent problems and opportunities, understand complex causal relationships, and continuously improve and adapt to changing conditions.

One of the most interesting aspects of big data, though perhaps disconcerting to some, is that it is not always necessary to understand causality. Insights can be gained by simply observing correlations gleaned from massive quantities of raw data gathered from diverse sources within large, complex systems, and useful patterns of behavior may appear even when causality is not understood.

With big data we can observe hidden patterns of consumer behavior, but talking to people — understanding how something positively or negatively impacts their experiences and their lives — affects us in an entirely different way.

An insight is useful, however, only if we can do something with it. Absent context and meaning, knowledge is simply interesting but not useful. In our view, we cannot reliably compute context and meaning when purpose isn't clear. Human insight is necessary to frame a situation properly, interpret the analysis, and choose how to move forward with the design of experiments and adoption of improvement and innovation ideas.

To this end, pioneering data science professors at Harvard University propose a simple five-step data science process:[12]

1. Ask a question
2. Get the data
3. Explore the data
4. Model the data
5. Communicate the data

"While computers are getting better and better at doing steps 2, 3, and 4," says Mark Basalla, lead data scientist at USAA, a financial services company, "only humans can ask the right questions in step 1 and tell the right story in step 5 in order to enable the right decisions to be made. Lean Thinking is an ideal way to help people do this consistently and effectively, through a variety of simple but effective problem-solving techniques and behaviors."[13]

In our experience with Lean transformation of large organizations, working with large cross-functional teams of human beings in complex and ambiguous circumstances, the question of purpose is often nuanced, involving the values and prevailing culture of the organization. A deliberate inquiry into purpose never fails to unlock deeper conversations, and it often leads to new understanding, creating energized, purpose-driven teams that apply a very different mindset to the problem or opportunity at hand.

What about empathy, for the customer and those doing the work to serve them? We must intentionally nurture the subtle, holistic awareness that we get as we go to gemba and observe the entire situation with all of our senses. With big data we can observe hidden patterns of consumer behavior, but talking to people — understanding how something positively or negatively impacts their experiences and their lives — affects us in an entirely different way.

And what about true, outside-the-box innovation? Some of the greatest leaps in human achievement have come when someone, in a moment of inspiration, looks at a situation in an entirely new context. Steve Jobs once said, "When you ask creative people how they did something, they feel a little guilty because they didn't really do it, they just saw something."[14] How do we compute that? Shunryu Suzuki, founder of the San Francisco Zen Center, sums up the challenge: "In the beginner's mind there are many possibilities, but in the expert's there are few."[15] The same, perhaps, can be said for an expert system.

In this article, we do not present an argument against big data, cognitive computing, and whatever lies beyond. Clearly we must learn to improve the human-computer symbiosis if we are to harness this qualitative leap in understanding and improving the world around us. Rather, we argue for maintaining a healthy respect for the role of human capability and insight, recognizing that decision making must be a combination of technical and social aptitudes. Only by understanding the

purpose of the organization, and the value it delivers to its customers both now and in the future, can we utilize this new technology properly. Both raw analysis and human creativity are necessary. They must be kept in balance, acting as catalysts for experimentation, continuous improvement, and innovation — the essence of Lean Thinking.

To make the right decisions, to ensure that we're optimizing for the appropriate outcomes, we must look deep inside the situation, ask the right questions, infer the causes for the correlations, design the right experiments, solve the right problems, think outside the box, and develop empathy for our customer, while always reflecting on our purpose. We must ensure that, when engaging with big data, the guiding hand of human understanding, intuition, and empathy is always present in the management systems and the culture of the organization.

Acknowledgment

We deeply appreciate the efforts of a coauthor who must remain anonymous. His insights as a data scientist, Six Sigma Master Black Belt, PMP, and mechanical engineer, and his practice in the aerospace, nuclear, chemical, financial, and IT domains, have been very helpful.

Endnotes

[1]Isaacson, Walter. *The Innovators: How a Group of Hackers, Geniuses, and Geeks Created the Digital Revolution*. Simon & Schuster, 2014.

[2]Isaacson (see 1).

[3]Kelly, John E., III, and Steve Hamm. *Smart Machines: IBM's Watson and the Era of Cognitive Computing*. Columbia University Press, 2013.

[4]Liker, Jeffrey K. *The Toyota Way: 14 Management Principles from the World's Greatest Manufacturer*. McGraw-Hill Education, 2004.

[5]Womack, Jim. "The Mind of the Lean Manager." Lean Enterprise Institute, 30 July 2009 (www.lean.org/womack/DisplayObject.cfm?o=1083).

[6]Deming, W. Edwards. *Out of the Crisis*. MIT Press, 1982.

[7]De Geus, Arie. "Planning as Learning." *Harvard Business Review*, March 1988 (https://hbr.org/1988/03/planning-as-learning).

[8]Womack, Jim. "Purpose, Process, People." Lean Enterprise Institute, 12 June 2006 (www.lean.org/womack/DisplayObject.cfm?o=742).

[9]This image was created by David Verble, former Toyota change agent and coach, and Lean Enterprise Institute faculty member, based on the original Purpose, Process, People work by Jim Womack and its later elaboration by Dan Jones.

[10]Deming (see 6).

[11]These steps are consistent with two popular continuous improvement cycles: Deming/Shewhart Plan-Do-Check-Act (PDCA) and Six Sigma Define, Measure, Analyze, Improve and Control (DMAIC).

[12]Mayo, Matthew. "The Data Science Process, Rediscovered." KDnuggets, 23 March 2016 (www.kdnuggets.com/2016/03/data-science-process-rediscovered.html).

[13]Basalla, Mark. Personal communication.

[14]Claburn, Thomas. "Steve Jobs: 11 Acts of Vision." *InformationWeek*, 7 October 2011 (www.informationweek.com/it-leadership/steve-jobs-11-acts-of-vision/d/d-id/1100596?).

[15]Suzuki, Shunryu. *Zen Mind, Beginner's Mind*. Shambhala Library, 2006.

Steve Bell is a faculty member of the Lean Enterprise Institute and the Lean Global Network, cofounder of Lean IT Strategies, winner of the Shingo Research Prize, and author of Lean IT *and* Run Grow Transform. *He can be reached at steveb@leanitstrategies.com.*

Karen Whitley Bell is cofounder of Lean IT Strategies, a 20-year healthcare veteran, and an award-winning author with a passion for research and human capability development. She can be reached at karenw@leanitstrategies.com.

A Strategic Approach to Big Data: Key to Analytical Success

by Bhuvan Unhelkar

Beyond Keywords

The term "big data" encompasses a wide variety of topics led by the two keywords "analytics" and "technologies." Technically, big data implies Hadoop/HDFS, Spark, and, at the back end, NoSQL. From a business viewpoint, however, big data analytics command greater interest as they enable identification of patterns, facilitate predictions, and also provide prescriptive advice for better decision making.

When medium to large enterprises want to adopt big data, they need to go through the rigors of large-scale adoption through people, processes, and technologies. For example, Guest Editor Barry Devlin's 4 Ps (Preparation, People, Prediction, and Production) provide one such basis for analytics adoption. However, analytics need to be coupled with a proper understanding of technological capabilities provided by enterprise architectures. We thus find that multiple technical, analytical, and architectural elements come into play in big data adoption.

This article argues for an overarching framework that will not only facilitate adoption of analytics and technologies, but will also provide a solid foundation for taking a strategic approach to big data. This framework is called the Big Data Framework for Agile Business (BDFAB v1.5), and its development is based on a review of the relevant literature, experimentation, and practical application. The key elements composing this framework are:

- Agile values for business
- Organizational roles in big data
- Building blocks of big data strategies for business (including the role of analytics within those strategies)
- Key artifacts in big data adoption
- Business conditions and limitations
- Agile practices
- A compendium (repository)

The building blocks of big data strategies are themselves are made up of five modules:

1. Business decisions
2. Data: technology and analytics
3. User experience: operational excellence
4. Quality dimensions
5. People (capabilities)

In addition, this framework is accompanied by a 12-lane process for big data transformation, especially in large organizations. Exploration of BDFAB will be of practical benefit to organizations looking for a sensible pathway into big data. At the same time, the framework provides opportunity for refinement based on further experimentation.

BDFAB Overview

As noted above, BDFAB is a research-based framework that facilitates a strategic approach to the application of big data to business. Most contemporary big data approaches focus either on the Hadoop ecosystem (as a suite of technologies, programming, and management) or on the analytics (based around extensive statistical techniques such as predictive analytics, net promoter score [NPS], and so on). This represents a significant lacuna in the big data space, which requires a comprehensive and holistic approach to formulating a business strategy and synergizing the aforementioned technical and analytical elements. This lack is also felt in the Agile space, which predominantly constitutes a methodical approach to solutions development.

Elsewhere I have argued that Agile needs to transcend the solutions space and move into business strategy.[1] The business technology domain finds itself awash in data and technology that can potentially be used to render a business Agile. Such strategic utilization of data requires deeper understanding of the current state of the business, its directions, and its capabilities (both architectural and people), as well as dynamic, smarter risk analysis. As the volume of data grows, the role of

information architecture is changing from the passive structuring and managing of data to a smarter, more active role in ensuring effective use of information.[2]

BDFAB builds on the technical and analytical aspects of big data in a holistic manner to understand and create new opportunities for business agility. Figure 1 highlights BDFAB's core philosophy of bringing together analytics and technologies but then going beyond them into the realms of agility and business strategy.

Wedding Big Data and Agility

In his foreword to *Big Data Analytics*, IBM Fellow and Chief Scientist Jeff Jonas observes how big data maps to agility in business [emphasis mine]:[3]

- "Organizations must be able to *sense* and *respond* to transactions happening now." (Agility is the ability to spot the changes coming through — which are transactions at both the micro and macro levels.)
- They also "must be able to *deeply reflect* on what has been observed — this deep reflection is a necessary activity to discover relevant weak signal and emerging patterns." (Agility requires the ability to take effective decisions; this effectiveness results from deep reflection, aided and impacted by big data analytics.)
- "As the feedback loop gets faster and tighter, it significantly enhances the discovery [from deep reflection]." (Agility requires rapid response, which in turn is based on analytical insights and leanness of organizational structure.)

Businesses can be helped to tap into Agile opportunities (ranging from expansion into new markets to enhancing customer satisfaction and/or optimizing internal business processes) by incorporating vital elements of Agile values, principles, and practices in big data adoption. The translation of these values from the depths of software development to business processes is the result of combining the formality of planned approaches and the flexibility of Agile, as in the Composite Agile Method and Strategy (CAMS).[4, 5]

BDFAB incorporates agility in a strategic business context with the understanding that Agile has transcended software development and now plays a major role in business organizations.[6] Agile is therefore a legitimate business goal in its own right,[7] and a strategic approach to big data can go a long way in achieving that goal. Such an approach aims to make use of structured, semi-structured, and unstructured data, and the velocity and

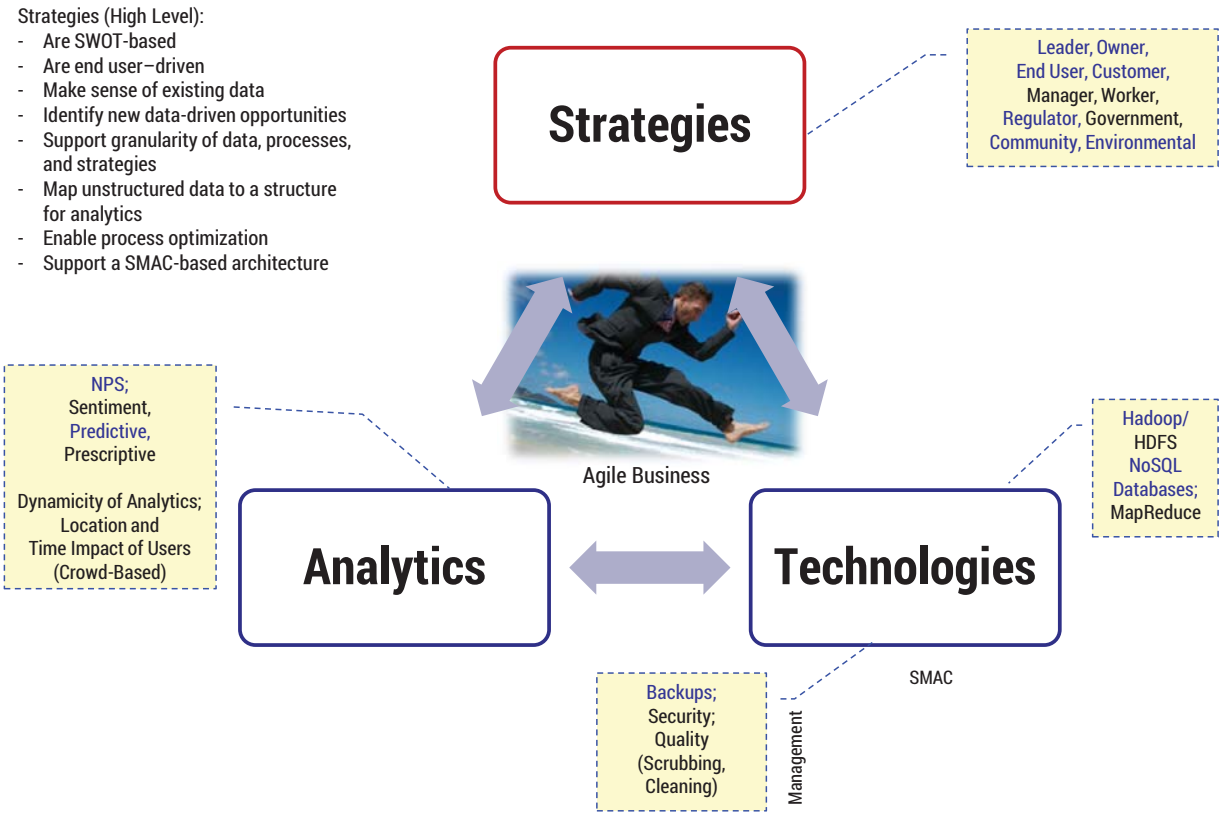


Figure 1 — Positioning big data strategies: transcending analytics and technologies. (Source: MethodScience.)

volume of such data to generate an ongoing and significant amount of business intelligence to enable improved business decision making.

Core Elements of BDFAB

Table 1 summarizes the core elements of BDFAB. The table also shows examples of BDFAB's core elements as well as key business considerations in utilizing a particular core element.

BDFAB starts with the business organization itself — its strengths, weaknesses, opportunities, and threats (SWOT) — and then moves into clarifying the vision and the capabilities the organization needs to satisfy that vision. Helping the business identify and exploit the existing and growing data capabilities (technologies and analytics) with a continuous focus on business decision making results in relevant insights. The output of this activity enables decision makers to either modify or develop new products and services, respond to individual customer issues, and rapidly change the business processes.

BDFAB helps position the technologies of Hadoop/MapReduce/Spark in a way that will assist in enhancing business agility. The decision makers need to see the correlation between big data (and its analytics) and making the business "Agile" as a result of those analytics. For example, online sellers (e.g., Amazon) use statistical analysis of hourly sales of books to recommend additional titles to readers. Airlines use hourly flight booking data to mark up (or down) their fares in a very Agile manner. Analytics thus continue to apply statistical techniques to generate rapid insights. Big data technologies, based around the Hadoop ecosystem (including HDFS, NoSQL, and MapReduce), support such rapid analytics by storing, sharing, and processing vast amounts of structured and unstructured data.

Furthermore, the analytics themselves are no longer static; they are themselves changing depending on the circumstances of an individual customer and/or the context in which the business finds itself (e.g., political uncertainty, changing legal structure, global collaborations). Thus, not only do the analytic processes support business agility, these analytical processes themselves need to be agile (or, in other words, continuously changing). BDFAB uplifts the capabilities of decision makers on an ongoing basis, resulting in business agility in a unique and holistic way.

The key innovative aspect of BDFAB is its focus on business strategies resulting from a balanced combination of big data technologies and analytics together with the concepts of composite Agile. This incorporation of composite Agile (CAMS) in BDFAB is based on the premise that Agile has transcended software development and now plays a major role in processes associated with the business.

BDFAB Modules

While analytics — including OLAP cubes, text and data mining, and dashboards — all add to and aid in decision making,[8] what is even more interesting is the

Core Elements of Framework	Examples of Each Core Element	Business Considerations
Values	Agility, insights, collaborations	What does the business aspire to (to-be state)?
Roles (people)	Data scientist, user, analyst, coach, investor	Who are the people to make it happen? To benefit?
Building blocks (phases)	Business decision (SWOT), technology, user experience, quality, people	Why do it (business reasons)? How to do it (phases)?
Artifacts	Plans (financial, ROI), feedback, approach, staff, center of excellence	What to produce? To use?
Conditions	Type, size of business (as-is)	Where and when to apply BDFAB?
Practices (Agile)	Stand-ups, stories, showcase	How to undertake agility at the change level?
Compendium (repository)	Manifesto, strategy cube, adoption process	How to guide change management, transformation?

Table 1 — Overview of BDFAB v1.5. (Source: MethodScience.)

strategy for putting this whole process together. How does one get an organization to reach a stage where these analytics and the decision making they enable become the norm? To achieve that goal, BDFAB comprises five major modules, as summarized in Table 2.

Business Decisions

This module introduces the concept of big data to the business and starts positioning big data as a basis for business strategies. Therefore, this module focuses on the strategic/business value of big data analytics. It begins with a SWOT analysis and moves into the risks, cost advantages, and adoption approaches to big data. ROI (and cost-benefit analysis) for big data adoption forms part of this module. Agility is introduced to the organization as a business value (transcending Agile used in software development projects).

Data: Technology and Analytics

The second module focuses on data analytics, mapping volume, variety, and velocity with structured, unstructured, and semistructured data types. Each of these characteristics of big data is invaluable in supporting corresponding business strategies — if properly formulated. This module demonstrates the interplay between analysis of data and its impact on creating business strategies. This module helps the business understand and incorporate structured, semistructured and unstructured data in its analytics and decision-making processes. It is underpinned by the technologies of Hadoop and the associated technical ecosystem.

User Experience: Operational Excellence

This module explains how data analytics can render a business Agile. Understanding customer (user) sentiments through a user experience analysis framework (UXAF) is the starting point for this work. Most UXAFs focus on time T0 to time T1 — when the user is in direct contact with the business through its systems and interfaces. Substantial data is generated, however, by user interactions with social and mobile networks that occur before T0 and after T1. Exploring the generation and use of this data (based around the SMAC stack) is part of the discussion in this module. The "predictive" and "prescriptive" nature of ensuing analytics is discussed here.

Quality Dimensions and the SMAC Stack

Quality considerations in the big data domain assume prominence because of the direct impact they have on business decision making. This module focuses on this crucial quality aspect in big data solutions: data, information, analytics (intelligence), processes, usability, and reliability. Uniqueness of unstructured data and what can be done to enhance and validate its quality are part of this discussion. The challenges of contemporary testing (and the role of Agile practices, such as continuous testing) together with their application to big data are also explained and implemented in business through this module.

People (Capability)

Cutter Senior Consultant Larissa Moss and data strategies expert Sid Adelman have discussed the ever-growing importance of people and their capabilities in the big data space.[9] Similarly, a McKinsey report on big data goes into the details of existing and needed capabilities in the big data technologies and analytics domains.[10] BDFAB incorporates this important people issue in big data adoption by identifying and enhancing the people capabilities at both the technical and analytical levels. The Skills

Modules	Brief Description
Business decisions	Focuses on existing capabilities, future vision, and a SWOT analysis.
Data: technology and analytics	Understands and builds on the technical capabilities of the Hadoop ecosystem and the volume, variety, and velocity of big data.
User experience: operational excellence	Builds on the value theme; analytics are performed before the user comes into contact with the business and continue well after that.
Quality dimensions	Examines the technical, economic, social, and process dimensions of a business, which are affected by big data. Social-Mobile-Analytics-Cloud (the SMAC stack) are also examined here.
People (capabilities)	Uses the Skills Framework for the Information Age (SFIA) to uplift the capabilities of the organization's people in the context of big data.

Table 2 — The five major modules of BDFAB v1.5. (Source: MethodScience.)

Framework for the Information Age (SFIA) is used as a backdrop for ascertaining current skill levels and mapping an up-skilling path for the human resources in organizations adopting big data. This module can lead to formation of centers of excellence around big data and related disciplines.

BDFAB in Practice

When an enterprise adopts big data using the BDFAB, it reduces its risks and gains the following practical advantages.

Creating a Business Advantage

BDFAB is meant to help a business adopt big data in a way that will result in business agility. The advantage of this framework results from bringing together two important concepts of modern-day technology and business — agility and big data. This synergy demonstrates the value of analytics in rapid business decision making.

CAMS embodies Agile characteristics (e.g., flexible, change-welcoming, iterative, collaborative, and ready to fail fast) that provide value to business.[11] BDFAB expands on these concepts to provide the business with the values of Insight, Collaboration, Dynamicity, Leanness, Governance, and Environment (i.e., sustainability).

Adoption Process

A 12-lane adoption process (see Figure 2) is a crucial part of BDFAB. This detailed adoption process provides guidelines in terms of which aspects of agility and big data should be adopted first, how the adoption should iterate, and how to identify and overcome blockers. It offers a basis for risk reduction in big data adoption.

Risk Analysis

BDFAB addresses risk in two ways. First, and most important, is the embedding of business risk considerations in the framework itself. BDFAB starts with a SWOT analysis that enables assessment of early business decisions relating to big data. The adopting

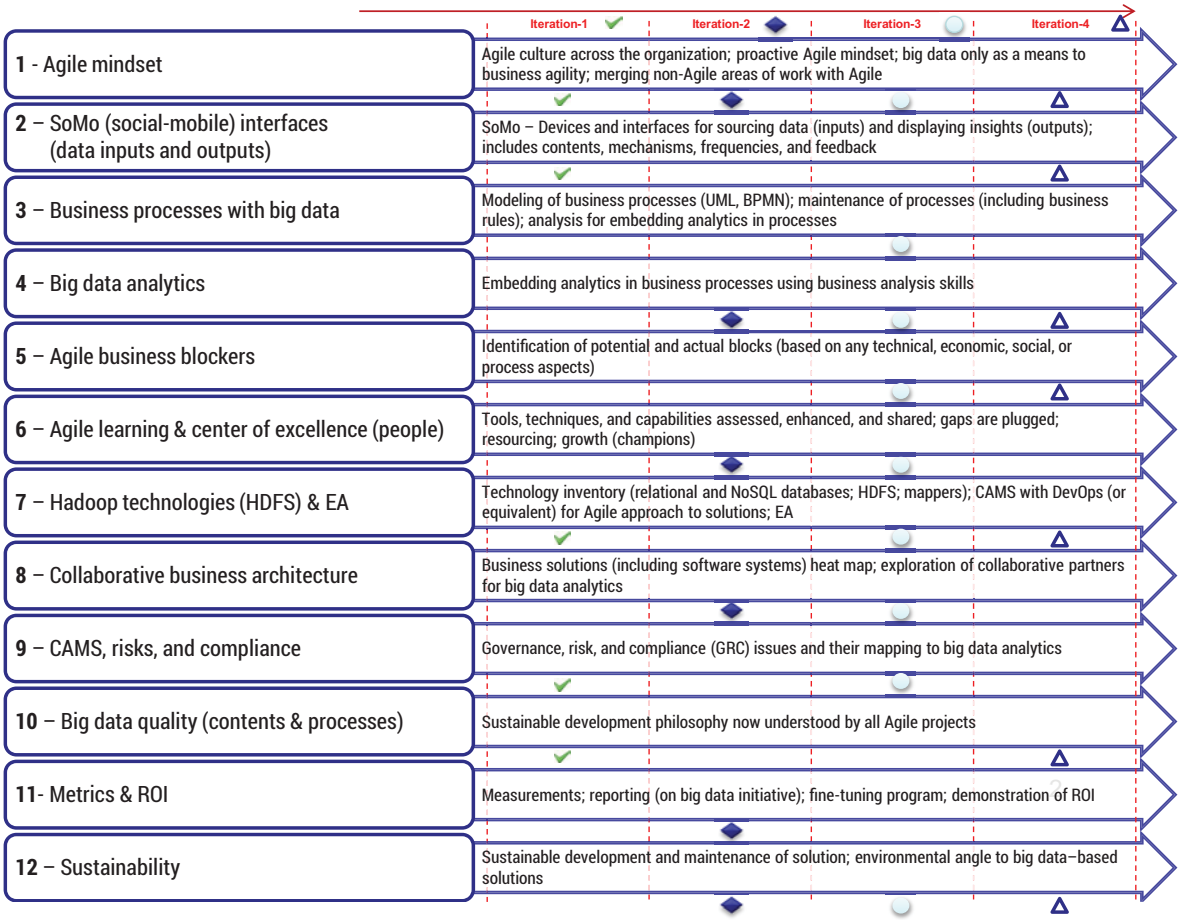


Figure 2 — Big data–driven Agile business: adoption roadmap. (Source: MethodScience.)

organization considers the business priorities, risks, and budgets in detail when undertaking this risk analysis. Furthermore, BDFAB provides multiple angles to big data adoption (including technologies, analytics, user experience, and people up-skilling), which results in risk reduction. Risk analysis also includes due consideration of the needs of different types and sizes of businesses. For example, a medium-sized travel business may be more interested in cloud-based analytics and not much up-skilling of staff, whereas a large bank will be interested in both cloud and people up-skilling. Having Agile concepts and values embedded in the framework is helpful in reducing business risks thanks to the rapidly iterative nature of agility.

Consumer Dialogue

Consumer dialogue and user experience are integral parts of BDFAB. The second module of BDFAB focuses on the user experience to ensure that it is included as a crucial element of the organization's big data strategy. For example, NPS statistical analysis provides a good basis for understanding what the consumer wants and embedding that in the business processes of the organization. Similarly, in optimizing internal business processes with big data–driven insights (e.g., anticipated production levels in a manufacturing plant, medical inventories in a hospital), BDFAB ensures continuous focus on the ongoing consumer dialogue.

Conclusion

In this article, I have presented BDFAB, a framework for adoption of big data by business. This comprehensive framework can help ensure that the end result of big data adoption is Agile business. BDFAB is unique in the sense that it elevates the current industry focus from technologies and analytics to business strategies. In due course, the framework will need to be accompanied by a corresponding process tool that will facilitate big data adoption in large organizations. Such a tool will not only help formulate organization-specific strategy, but also enable monitoring, tracking, reporting on, and optimizing the process of big data adoption for business agility.

Endnotes

[1]Unhelkar, Bhuvan. *The Art of Agile Practice: A Composite Approach for Projects and Organizations*. CRC Press, Auerbach Publications, 2013.

[2]Evernden, Roger. "Information Architecture: Dealing with Too Much Data." Cutter Consortium Business & Enterprise Architecture *Executive Report*, October 2012.

[3]Jonas, Jeff. Foreword to *Big Data Analytics: Disruptive Technologies for Changing the Game*, by Arvind Sathi. MC Press Online, 2012.

[4]Unhelkar (see 1).

[5]Mistry, Nosh, and Bhuvan Unhelkar. "Composite Agile Method and Strategy: A Balancing Act." Paper presented at the *Agile Testing Leadership Conference 2015*, Sydney, Australia, August 2015.

[6]Unhelkar, Bhuvan. "Agile in Practice: A Composite Approach." Cutter Consortium Agile Project Management & Software Engineering Excellence *Executive Report*, Vol. 11, No. 1, 2010.

[7]Unhelkar (see 1).

[8]Kudyba, Stephan. *Big Data, Mining, and Analytics: Components of Strategic Decision Making*. CRC Press, Auerbach Publications, 2014.

[9]Moss, Larissa T., and Sid Adelman. "The Role of Chief Data Officer in the 21st Century." Cutter Consortium Data Analytics & Digital Technologies *Executive Repor*t, Vol. 13, No. 2, 2013.

[10]Manyika, James, et al. "Big Data: The Next Frontier for Innovation, Competition, and Productivity." McKinsey Global Institute, May 2011 (www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation).

[11]This is discussed in detail in: Unhelkar, Bhuvan. "The Psychology of Agile: Fundamentals Beyond the Manifesto." Cutter Consortium Agile Project Management & Software Engineering Excellence *Executive Report*, Vol. 14, No. 5, 2013; and Unhelkar, Bhuvan. "The Psychology of Agile: Group Dynamics and Organizational Adoption." Cutter Consortium Agile Project Management & Software Engineering Excellence *Executive Report*, Vol. 16, No. 4, 2015.

Bhuvan Unhelkar (BE, MDBA, MSc, PhD; FACS) is a Senior Consultant with Cutter Consortium's Business Technology & Digital Transformation Strategies practice. He has more than two decades of strategic, as well as hands-on, professional experience in the information and communication technologies (ICT) industry. He is Associate Professor and Chair of the IT Department at the University of South Florida (Sarasota-Manatee). As a Founder of MethodScience.com, he has demonstrated consulting and training expertise in big data (strategies), business analysis (use cases, BPMN), software engineering (object modeling, Agile processes, and quality), collaborative Web services, green IT (environment), and mobile business. His domain experience includes banking, financial, insurance, government, and telecommunications organizations. Dr. Unhelkar earned his PhD in the area of object orientation from the University of Technology, Sydney, in 1997. Since then, he has authored/edited 17 books in the areas of collaborative business, globalization, mobile business, software quality, business analysis/processes, the UML, and green ICT, and he has extensively presented and published papers and case studies. Dr. Unhelkar is a sought-after speaker, a Fellow of the Australian Computer Society (elected to this prestigious membership grade in 2002 for distinguished contribution to the field of ICT), Life Member of Computer Society of India and BMA, Rotarian (Paul Harris Fellow), Discovery Volunteer at NSW parks and wildlife, and a previous TiE Mentor. He can be reached at bunhelkar@cutter.com.

Maximizing Analytic Value: Attributes a NoSQL Analytics System Must Have

by Jeff Carr

Overview

Semistructured data, called "NoSQL" data in this article,[1] is growing at an unprecedented rate. This growth is fueled, in part, by the proliferation of Web and mobile applications, APIs, event-oriented data, sensor data, machine learning, and the Internet of Things, all of which are disproportionately powered by NoSQL technologies and data models.

NoSQL operational databases continue to gain ground on relational databases, with MongoDB recently becoming the fourth most popular database in the world. Hadoop is well on its way to becoming the de facto "data lake" for company-wide data, regardless of structure, and the rapid maturation of machine learning has provided robust ways to turn unstructured data like video, audio, and images into NoSQL data.

Recently, there has been a flurry of discussion about the implications of the rise of NoSQL for analytics and decision support systems. These discussions often revolve around use cases, the extent to which nontabular data models permit analytics (and if so what kind), and whether or not NoSQL systems have the ability to participate in analytic workloads. As expected for any early-stage technology, these discussions are often imprecise and conflate a wide range of concerns, including semantics, architecture, performance, technology, use cases, and user interface.

This article explores a single concern: describing the system-level capabilities required to derive maximum analytic value from a generalized model of NoSQL data. A generalized model is a model that works across all data sources no matter what type of data is present. Generalized analytics can answer all questions, from simple to complex, across all data types. This approach leads to eight well-defined, objective attributes, which collectively form a precise capabilities-based definition of a NoSQL analytics system.

These capabilities are inextricably motivated by use cases, but other considerations are explicitly ignored. They are ignored not because they are unimportant (quite the contrary), but because they are orthogonal to the raw capabilities a system must possess to be capable of deriving analytic value from NoSQL data.

Tools that have these eight capabilities will set in motion a large shift and almost complete recalibration of technology strategy and architecture, fomenting a new era of application development and innovation. Put another way, our current technology is stitched together by RDBMS technology. When we can finally add NoSQL analytics to the NoSQL data sources, we will effectively have the two necessary ingredients to deliver complete business value for the post-relational age.

The Nature of NoSQL Data

To discuss a NoSQL analytics system, we must first have a coherent definition of the term NoSQL data, as well as some assurance that this definition permits enough abstraction to formally model general-purpose analytical capabilities. In the relational data model, the core abstraction for data is a structurally homogeneous set of tuples of atomic values. NoSQL generalizes this to arbitrary data structures of the type found in programming languages like Javascript.

NoSQL analytics, then, refers to analytics over arbitrary data structures. While a system capable of extracting analytic value over arbitrary data structures might sound intractably complex, the variation can be abstracted with a few building blocks. To motivate this abstraction, and establish the type of data that a NoSQL analytics system is expected to support, the following sections review the primary sources of NoSQL data and the data models that they employ.

APIs

There are well over 200,000 APIs in the world. Without exception, each of these APIs accepts and produces NoSQL data. The concept of producing relational data from an API does not even make sense, as it would

require an API that could accept and produce a database. As APIs increasingly become the fabric that binds technology together, they will continue to be an inexhaustible source of NoSQL data.

APIs are not databases, but they do invariably expose database-like mechanisms for querying — including filtering, shaping, and in some cases aggregation. API data is also frequently stored in files, databases, and data lakes for subsequent analysis. The primary data formats employed by APIs are as follows:

- **JSON** — an acronym for JavaScript Object Notation, JSON is by far the most common API format, thanks to the simplicity with which it can be generated and parsed.
- **XML** — an acronym for Extensible Markup Language, XML still plays a role in many APIs, particularly SOAP, and its close relative, HTML, is the primary medium of content for the Web.

NoSQL Databases

NoSQL databases have existed since the 1960s but have only proliferated in number and surged in popularity in the past decade. Originally driven solely by the need for Web-scale storage, today companies adopt NoSQL databases only partially for horizontal scalability.

NoSQL databases provide an operational ease rarely seen with RDBMSs. In addition, they provide agility and flexibility not possible in the relational model and an order-of-magnitude performance improvement for certain classes of problems. Finally, because they support much richer data models than relational systems, they make it possible to build more complex applications with substantially less effort.

There is little standardization among NoSQL databases. Every database exposes its own unique set of APIs, its own data model, and its own query language (if distinct from the APIs). Despite this heterogeneity, NoSQL databases can be classified into categories based on the type of data model they support. These categories include key/value-oriented, document-oriented, graph-oriented, data structure–oriented, and wide column–oriented (among others). The data models of a few common NoSQL databases are presented below:

- **MongoDB** — a document-oriented database that supports a strict superset of JSON, including arbitrary nesting of subdocuments and arrays, as well as leaf types such as integers, floating-point numbers, strings, and date/times
- **Aerospike** — a data structure–oriented database that supports arbitrary nesting of lists and maps, as well as leaf types such as integers and strings
- **Redis** — a data structure–oriented database that supports flat lists, sets, maps, and string leaf types (in practice, these strings often store data structures such as JSON)
- **CouchDB** — a document-oriented database that supports arbitrary JSON
- **ElasticSearch** — a document-oriented database that supports arbitrary JSON
- **MarkLogic** — a document-oriented database that supports arbitrary XML, as well as JSON via a conversion layer
- **Clusterpoint** — a document-oriented database that supports hierarchical documents that can encode JSON, XML, and similar content
- **Neo4j** — a graph-oriented database that supports values that contain typed references (edges) to other values and a mapping from string to values (properties), such as numbers, booleans, strings, and arrays of the above

As APIs increasingly become the fabric that binds technology together, they will continue to be an inexhaustible source of NoSQL data.

Big Data

Hadoop has made commonplace the notions of "infinite" file systems and localized data computation. This, in turn, has made it increasingly common to store, archive, and process massive quantities of data in Hadoop. Some common file formats for Hadoop include JSON, XML, ORC, Avro, and Parquet, all of which support storage of denormalized data (some heterogeneous, some homogeneous). These data formats are self-describing and self-contained, so a single file can contain a complete description of any kind of non-cyclic data. As a result, a large percentage of the data in big data file systems is stored in such formats.

A Generic Data Model for NoSQL

As we can see from the preceding review, NoSQL is an amalgamation of everything non- and post-relational.

Instead of standardization and uniformity, the moniker represents a multitude of databases, data models, and data formats. In practice, however, a few building blocks are sufficient to represent nearly all NoSQL data:

- **A heterogeneous ordered map from value to value.** When the keys are strings, this is often called a record, object, or document in NoSQL systems. In the general case, however, the keys need not be strings and can themselves be arbitrary values.
  - While many systems do not care about or provide ordering, some do, so the more general notion is an ordered map (i.e., a map whose key-value pairs have some user-defined ordering).
  - Maps can also represent sets, as a mapping from a unique identifier to a value.
  - Neither the keys nor the value need have the same type, which allows a direct encoding of heterogeneity.

NoSQL is an amalgamation of everything non- and post-relational. Instead of standardization and uniformity, the moniker represents a multitude of databases, data models, and data formats.

- **A heterogeneous ordered array of values.** Unlike arrays in relational systems (which are poorly supported and not used much), arrays in NoSQL systems occur frequently and can contain arbitrary values of completely different types.
  - Ordered arrays can also be used to represent unordered collections.
- **A value reference.** A reference is a link to another value. This is called a "foreign key" in relational systems, an "edge" in NoSQL graph systems, and a "reference" in most programming data models.
- **Atomic values.** Atomic values do not contain any other value; they are the primitive types of a data model. Usually, they include things like booleans, numbers, characters, dates, times, date/times, intervals, and so forth.
  - Text is actually not atomic, as it can be represented using arrays of characters.

Any NoSQL analytics system that abstracts across different NoSQL data models will inevitably end up using a generalization that is similar (if not identical) to this one.

Approaches to NoSQL Analytics

In the history of NoSQL data, there have been many approaches to solving the problem of performing analytics on it. Not all of these approaches are able to solve all problems in NoSQL analytics — they vary greatly in their expressiveness and flexibility. In this section, I will survey some of these approaches and conclude by highlighting some of the recent work in open source aimed at making NoSQL analytics truly first-class.

Coding and ETL

NoSQL storage systems first arose in the 1960s. Despite the existence of data interface languages in products like IBM's IMS (a hierarchical database built in 1966), analytics on non-relational data has historically required one of two approaches:

1. **Custom coding.** Data is pulled out of a NoSQL source and filtered, shaped, and aggregated by hand-written code. This approach is common today with NoSQL operational databases, especially in smaller companies with less sophisticated analytical needs.
2. **ETL.** Data is pulled out of a NoSQL source, transformed, and flattened to a simpler relational data model. This approach is also common today, primarily among larger companies that have advanced analytical needs, a scarcity of development resources, and heavy investment in legacy relational analytics tooling.

Hadoop

The rise of Hadoop made semistructured data much more common. This, in turn, created the need for analytical capabilities on semistructured data. As a general-purpose computing platform, Hadoop's Map/Reduce framework has supported near arbitrary analytics on NoSQL data from the beginning. Originally, however, these capabilities could be leveraged only by skilled big data engineers.

This engineering burden led to the creation of Pig and Hive, two complementary (but overlapping) technologies that support basic analytics over semistructured data. Pig adopted an expressive NoSQL data model supporting bags, tuples, maps, and more, and exposed functionality sufficient for many common analytic scenarios. For more advanced analytic scenarios, Pig

supported a pluggable UDF architecture. Hive, meanwhile, provided a simple, if first-class, concession to nested data in the form of lateral views, a feature that — while unfamiliar to those coming from a relational background — proved indispensable to the highly nested world of NoSQL data. Other technologies in use for analytics on semistructured data in Hadoop include Spark, Cascading, and other computational frameworks that rely on hand-written code.

Real-Time Analytics

Many NoSQL operational databases have acquired the ability to perform atomic operations, such as increment and decrement on numeric values. Combined with dynamic data models, this allows NoSQL systems to perform so-called real-time analytics, in which various aggregations are built dynamically. While the flexibility of such real-time analytical systems is poor, they scale easily and provide a basic level of insight into simple event-oriented systems. The analytics produced by such systems are already aggregated, but further filtering and aggregation are possible, so real-time analytics is at best a partial solution to the problem of NoSQL analytics.

Relational Model Virtualization

As NoSQL data systems have slowly entered the mainstream, there has been growing demand for the ability to connect relational analytics tools (such as Tableau, Qlik, Cognos, MicroStrategy, and BusinessObjects) to these NoSQL systems. This has spawned so-called relational model virtualization adapters. Usually employing the JDBC or ODBC protocols, these adapters expose a virtual relational model on top of a NoSQL data system. The most sophisticated of these drivers expose virtual tables for arrays and data nesting and use null-padding to encode heterogeneity.

These drivers are not without application, but customer satisfaction has been poor — partially because of performance issues, but mostly because of the impedance mismatch between relational and NoSQL data models. As NoSQL analytics systems develop, it will become apparent that virtualization suffers from an inability to answer many types of analytic questions over NoSQL data.

First-Class NoSQL Analytics

Recently, the industry has entered a new era for NoSQL analytics. The need for analytical capabilities over semistructured data no longer requires justification. Instead, these needs are assumed, and the only point of contention is the expressiveness of such capabilities.

In the past few years, many relational systems have added one or more new column types for semistructured data (typically JSON or XML). Some, such as PostgreSQL, allow indexing on the inner structure of this data. All expose basic capabilities for accessing such data, but the capabilities fall short of the eight attributes of NoSQL analytics systems.

Beyond these perfunctory concessions from relational systems, we have seen a new generation of analytics systems enter the scene, such as Drill, Quasar, and FORWARD. These systems were natively designed for performing analytics on semistructured data. Drill adopts a JSON-like data model, Quasar strives for full generality and expressiveness, and FORWARD lands somewhere in the middle. Though they differ on their approaches and level of expressiveness, all recognize the need for truly first-class NoSQL analytics.

On the standards front, SQL++ (FORWARD), SQL2 (Quasar), N1QL (Couchbase), and Impala's SQL extensions are among many efforts to generalize the relational query model to semistructured data. While a standard query interface has yet to emerge, the wealth of work being done in the space suggests that eventually the industry will see convergence, and that it will look a lot like SQL, but with a richer data model and multi-dimensional semantics.

Attributes of NoSQL Analytics Systems

The preceding sections have outlined the numerous approaches to the problem of NoSQL analytics. For each approach, there are many different solutions. Not all of these solutions are equal. In order to derive maximum analytic value from arbitrary NoSQL data, a solution must possess the following eight attributes:

1. Generic data model
2. Isomorphic data model
3. Multi-dimensionality
4. Unified schema/data
5. Post-relationality
6. Polymorphic queries
7. Dynamic type discovery and conversion
8. Structural patterns

These attributes may be used to judge whether or not any given system is capable of generalized NoSQL analytics as described in this article.

1. Generic Data Model

Attribute. NoSQL analytics systems must support a generic model of NoSQL that abstracts across the differences between different sources of NoSQL data.

To the extent that a NoSQL analytics system is truly general purpose, and capable of deriving analytic value from post-relational data models, it is necessary that the system be capable of working across complex NoSQL data, such as edges in a NoSQL graph database or maps with complex keys in a data structure–oriented database.

2. Isomorphic Data Model

Attribute. NoSQL analytics systems must support queries across the data as it is actually structured, or across an invertible view of the data that preserves all features of the original (and hence is isomorphic to the original data model).

NoSQL data models are rich, and the ways in which these models are used to capture and process information differs substantially from the relational world. Some systems adopt the strategy of exposing NoSQL data under a relational model; this fails, in part, because the virtual relational models do not contain the same information as the original data. They both lose information present in the original data and add other "fake" data, resulting in a poor approximation of the original. In order to preserve the maximum amount of analytic value from NoSQL data, NoSQL analytics systems must expose a completely lossless view of the original data. In the ideal scenario, a NoSQL analytics system exposes and allows analytics across the data as it is actually structured, with no changes to the rich data structures or heterogeneity present in the original data set.

Example

In the case of a content management system built on an operational NoSQL database, the data model may consist of individual pieces of content (represented as semistructured HTML) that have arrays of comments, each of which has information on the author of the comment. The content may also include a histogram of website visitors, broken down by day and browser type. A NoSQL analytics system should reflect and allow analytics on this structure exactly as it exists in the NoSQL database, or at minimum, reflect a view of this structure that preserves all information content of the original.

3. Multi-Dimensionality

Attribute. NoSQL analytics systems must support unrestricted lifting of set-level analytic operations to arbitrary dimensions of nested data.

The analytic utility of relational systems comes from their ability to perform set-level operations, such as filtering, grouping, and aggregation. In the relational world, the data model is always flat, and there exists a single dimension over which set-level operations may be applied: namely, the set of tuples under consideration. In contrast, NoSQL data is inherently multi-dimensional. These dimensions of data are nested and have irregular shapes. In order to derive analytic value from them, a NoSQL analytics system must allow the performance of all set-level operations on arbitrary dimensions of nested data.

Example

In the case of a behavioral analytics application built on an operational NoSQL database, the data model may consist of one value per user (which would contain an array of sessions). Inside each session might be an array of all events comprising the session. Events might have ad hoc structure (generated by Javascript or code running on smartphones) and may include further nesting, such as a sorted list of possible locations as derived by geo IP.

A NoSQL analytics system must allow arbitrary and unrestricted analytics on any of these nested dimensions of data. For example, the system must support returning a per-user histogram of events, broken down by hour of day, and also a per-province histogram of events, across all users, broken down by hour of day.

4. Unified Schema/Data

Attribute. NoSQL analytics systems must support the full range of analytic capabilities on the "schema," without any difference in analytic expressiveness between schema and value.

One of the most distinctive properties of NoSQL systems, which makes them strictly more powerful than their relational counterparts, is that the "schema" of NoSQL data is itself data. In fact, the very notion of schema breaks down in many NoSQL systems, because the schema refers to string keys in a map-like data structure. Although these keys may be used in a fashion similar to column names in a relational system, they may also be used for storing heterogeneous data, which has no direct parallel in a relational system. As a consequence, a NoSQL analytics system must support

completely arbitrary, ad hoc analytics on schema. A pleasing consequence of this support is that several operations that are classically impossible or extremely difficult to do in a relational system become trivially easy in a NoSQL analytics system (such as pivots).

Example

In the case of a real-time analytics application built on a NoSQL operational database, the keys in a map may represent date/times, while the values may be numbers that are incremented using atomic counters that are common to NoSQL databases. A NoSQL analytics system must be capable of pulling out the date/time values encoded in the schema, filtering them, joining them with other date/time histograms, and aggregating them across the joined data set for the same date/times.

5. Post-Relationality

Attribute. NoSQL analytics systems must be strictly more expressive than relational analytics systems.

In a NoSQL analytics system, the need for data denormalization is lessened, because NoSQL data models permit storing denormalized data directly. However, even with a high degree of denormalization, any given data set is still related to many others, and for analytic purposes, tying them together is essential. Thus, a NoSQL analytics system must be post-relational rather than non-relational, supporting the full expressive power of relational algebra, including joins, filters, groups, and aggregates.

Example

In the case of a dump of API responses for an online store, the data model might consist of heterogeneous product catalog data, each entry containing not only information on the product (which varies depending on whether the entry represents an event, subscription, product, or electronic product), but also user reviews and ratings. A NoSQL analytics system should be capable of joining the user review data, which is nested inside the product entries, to a user profile data set that maps from user IDs to profile information.

6. Polymorphic Queries

Attribute. NoSQL analytics systems must support queries across structurally polymorphic data.

In the limit, a collection of values may share absolutely no structure, with every value having a completely different structure from every other value. However, in many common cases, there are common structural elements across values that have the same semantics. NoSQL analytics systems must be able to query across such common structural elements of a collection of values, even when they possess arbitrarily large amounts of structural heterogeneity.

Example

In the case of a multi-tenant CRM application built on a NoSQL database, the data model for customer contacts may include common elements such as contact name and email, but may also include arbitrary user-defined data structures (possibly nested and heterogeneous depending on the contact type). A NoSQL analytics system must support queries across the common structural elements of these contacts despite the large degree of heterogeneity.

7. Dynamic Type Discovery and Conversion

Attribute. NoSQL analytics systems must support runtime type identification and conversion so that custom business logic can be used to dictate analytic treatment of variation.

Heterogeneity is a defining characteristic of NoSQL data. Values may have completely different structures. Elements that have the same semantics may have different structures, while sometimes elements that have the same structure may have different semantics. In order to enable a business to leverage its domain knowledge of the data, it is necessary for a NoSQL analytics system to expose the type (and therefore structure) of data at runtime and to enable conversion between different types according to custom logic. Most relational analytics systems already support type conversion, but NoSQL analytics systems must go beyond that to support type identification, as well as far richer conversions and identifications than would be necessary in a relational system, due to the richness of NoSQL data models.

Example

Over the lifetime of an application, a field in a record may have different types: for example, a comma-separated list of values embedded in a string or an array of values. A NoSQL analytics system must be capable of allowing queries to inspect the type/structure of the field and then to convert the structure as business logic dictates, splitting the string into an array based on the position of the commas.

8. Structural Patterns

Attribute. NoSQL analytics systems must support structural pattern matching that is capable of filtering

and extracting from variable-length, multi-dimensional patterns.

NoSQL data frequently represents content (such as Web pages, résumés, health forms, and so forth), events, and relationships. In such cases, many analytical use cases require the identification and extraction of user-defined patterns. Not that these use cases are all restricted to NoSQL data. Most relational analytics systems have a way to identify and extract user-defined patterns in event-oriented data (for example, MATCH in Vertica, NPATH in Aster, and the SQL window functions). Indeed, all relational systems can identify simple character patterns in strings with SQL's LIKE clause. For NoSQL data, these use cases are far more common, and they are substantially more complex because of the much richer data model.

Example

In the case of a static snapshot of the HTML for an entire website, the data model might consist of raw HTML pages. This data may be linked to transactional data. A NoSQL analytics system must support computing the purchase rate as a function of how far a purchase link is from its nearest (topside) header element. This requires the ability to match on a variable-length pattern consisting of a header, followed by zero or more intermediate nodes, followed by a block element that contains (at some unspecified but bounded depth) a link target that matches the pattern for a purchase link.

What's Possible Now?

We're at the beginning. There are a few companies in the vanguard of building NoSQL analytics systems. There are other companies that are allocating resources to test and prototype. Then there's the majority of companies that are on the sidelines: they have post-relational data, they know there are opportunities to extract value from the new technologies, but they don't have a process or tools. For this group, they are staring into the unknown. With the advent of NoSQL analytics, the missing piece turns a wobbly stone into a rock-solid foothold. It's a veritable stack now — or, a unit of production capacity. This article has precisely defined the capabilities required to extract maximum analytic value from arbitrary NoSQL data; that is, to precisely define what a NoSQL analytics system is capable of.

As a reminder, this challenge is complicated by the fact that there is no such thing as a single "NoSQL data model." However, the common NoSQL data models can all be unified with an abstract data model containing maps, arrays, references, and a variety of common atomic types. To derive maximum analytic value from NoSQL data, a system must:

1. Have a generic data model capable of abstracting across a wide range of NoSQL data models
2. Reflect back a lossless view of the data
3. Support set-level operations on arbitrary nested dimensions
4. Allow arbitrary analytics on schema
5. Combine new operators built for non-relational data and traditional relational operators
6. Allow queries across structurally polymorphic data
7. Enable dynamic type identification and conversion
8. Support multi-dimensional pattern matching

Together, these attributes form a robust, capabilities-based definition of what it means for a system to support generalized NoSQL analytics. Additionally, they serve as a guide for companies to evaluate competing approaches to NoSQL analytics. With this foundation of NoSQL analytics systems laid, there is broad room for exploration, differentiation, and innovation around other relevant dimensions, semantics, architecture, performance, technology, use cases, and user interface.

Welcome to the era of NoSQL analytics. It's just the beginning.

Endnote

[1]Technically, NoSQL refers to "Not Only SQL," and while the term has historically been used mainly to describe NoSQL operational databases, it applies equally to alternative data models.

Jeff Carr is cofounder and CEO of SlamData. Mr. Carr's career has focused on early stage companies in markets including network security, VOIP, and big data. He started his career with NCR Corporation, where he spent eight years in the Data Services Division until NCR's acquisition by AT&T. Mr. Carr then moved to Oracle in the applications division, focusing on major telcos in the western US and Latin America. In 1997 he joined Netscape Communications, where he led the central sales region team. Beginning in 2012, Mr. Carr served as COO of Precog, which built a SaaS data science platform. There he led sales, finance, and operations from pre-launch, through beta, and into revenue generation, which helped lead the company to acquisition by Rich Relevance in August 2013. Mr. Carr then founded SlamData to address the shortcomings of current legacy analytics solutions in the age of modern data and to address the inconsistency of APIs between NoSQL vendors. He can be reached at contact@slamdata.com.

Challenges to Maximizing the Value of Future Innovation in Big Data Analytics

by Donald E. Wynn, Jr., and Renée M.E. Pratt

Big data analytics (BDA) is arguably the hottest information technology phenomenon today. Large and small organizations, in virtually every industry, are enamored with the ability to gather and analyze what would have seemed to be an obscene amount of data only a few years ago. This ability has led to the generation of insights that would not have been possible due to the complexity of the underlying data. As a result, these technologies are reaching the point of being fully diffused throughout public and private organizations worldwide. Even with this broad diffusion, we have only scratched the surface of what organizations can accomplish with analytics. With the increasing proliferation of Internet of Things (IoT) devices, advances in data management technologies and analytical tools, and developments in algorithms, the opportunities to improve a wide range of organizational processes and outcomes are seemingly limitless.

For instance, healthcare organizations (HCOs) are using analytics to improve a number of financial, clinical, and operational objectives.[1] Ultimately, a fully mature adoption of analytics by HCOs will lead to personalized medicine, interventional decision support, and prescriptive analytical outputs, based on a broad array of structured and unstructured data inputs.[2] Another example is the automobile insurance industry, where BDA techniques allow firms to personalize policies based on the risk profile of individual drivers as measured by telemetry devices in the policyholder's car.[3]

But realization of this technologically utopian vision is years away. One observer notes that firms interested in developing useful analytics programs should expect that their first steps may in fact be "completely and utterly wrong."[4] As initial implementations have evolved, firms have seen difficulties arise in terms of strategic implementation, data management, and staffing concerns.

Challenges of BDA

Organizations seeking to incorporate effective analytics programs will likely encounter several challenges along the way. Whereas many of these can be dealt with in the short term, others will require solutions that we do not know to exist at the present time. In the balance of this article, we discuss several of the challenges and possible solutions, while addressing the components involved in any BDA plan.

Managing BDA Implementations

Implementing technologies that have a profound impact on your processes requires a consistent plan of attack over time. This plan should be expressed in terms of three key components of the information system (in addition to the technologies themselves):

1. **Processes** — the strategic and operational processes into which BDA are incorporated
2. **Data management** — a scheme for managing the necessary data
3. **Staffing** — the collective skills and expertise needed to pull these pieces together

The plan includes not only a picture of the current IT structure, but also a description of the expected benefits from the integrated platform that arises. This includes both growth strategies and integration planning. Organizations that are able to successfully manage the changes required in each of these components are better able to take advantage of the resultant capabilities.

Innovating according to these plans is an iterative process in which the results of previous implementations serve as a baseline from which new changes are developed. The changes in many ways resemble business process management, in which we begin with a

model of the existing system, identify the changes needed to arrive at a desired system, implement these changes, and evaluate the results against previous benchmarks and desired metrics (see Figure 1). The conclusion of the evaluation phase naturally leads to the contemplation of desired future changes.

The cycle begins with identifying what the ultimate system architecture should look like, in terms of each of the components and how they are integrated. Specifically, an organization should make some assumptions and targets for how BDA will eventually be incorporated in day-to day and strategic routines. For instance, it is expected that HCOs will eventually be able to capture information from patients' wearable devices, their own inputs, data from hospital stays and general practitioner visits, and a host of other information. Each HCO should therefore base its current procurement and implementation decisions on its expectations for future requirements for data storage, personnel capacities, and technological platforms.

However, not all organizations are in a race to deploy the latest technologies as soon as they hit the market. Some are more content to continue building on their existing systems until such time as they are capable of developing the insights needed to compete effectively against rivals using analytics. Of course, others are prone to innovate as soon as new technological advances come on the market. Regardless of the pace of innovation, the iterative process itself remains the same.

Integrating Innovation

As an organization progresses through several iterations of innovation, there are a number of concerns about integrating the changes with the existing information systems structure. Integrating changes in the technological platforms depends on the origin of the projects themselves. Generally speaking, the introduction of projects can be accomplished in either a top-down or bottom-up manner (see Table 1).

Top-Down

In organizations with relatively tight controls on their IT environments, analytic projects are often encouraged and managed from the top down. That is to say, they are introduced via formal project management strategies and techniques, with documented integration strategies. These organizations also tend to focus more on the existing legacy data and applications. Top-down is advisable for projects that have large-scale implementation requirements, shared data requirements, or significant strategic implications. Under these circumstances, the oversight and planning that come from such projects lead to a more organized result. However, this oversight frequently comes at the expense of speed to market and overall agility of the

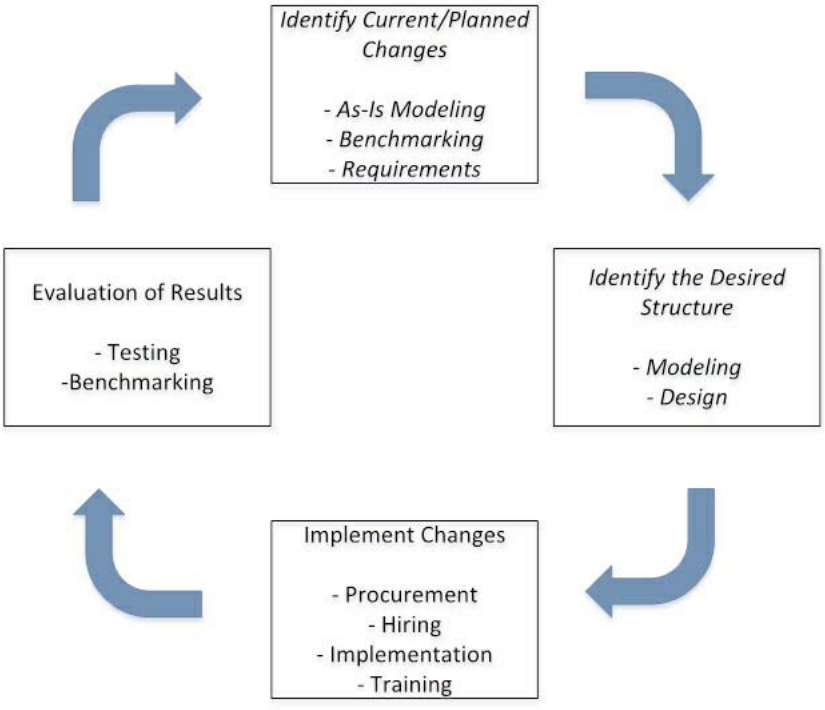


Figure 1 — Analytics process innovation.

IT function. It also adds to the initial costs for each iterative project.

Bottom-Up

To combat this expense and lack of agility, many firms encourage experimentation from the bottom up; that is, they allow individuals or groups to implement small-scale projects (such as analytics applications) for specific application uses. Typically, these solutions are designed to generate insights for specific departments or groups based on a limited amount of locally available data. As these smaller, targeted projects are developed, they are often not designed to be combined with legacy data sources or applications. Rather, the new applications are developed separately from the remaining data and applications, and even from each other. Eventually, as the small projects grow in scope or resource requirements, there is a need for more integration with the existing information systems architecture to avoid repetitively expensive rework and development to link these applications together.

Both top-down and bottom-up innovation processes are useful, albeit under different circumstances. For larger, more rigid organizational structures or large amounts of data trapped in legacy systems, the top-down approach is often a better fit between the organization and the IT department. For projects that are required by firms whose survival depends on increasingly nimble responses to industry forces, or those projects whose scope is limited to specific departments and somewhat isolated data requirements, a bottom-up approach is likely to be more beneficial.

Perhaps the best approach for many organizations resembles a bimodal IT structure in which smaller, focused implementations are developed for specific applications, while the overall enterprise structure remains focused on the goals of the entire organization. But this ultimately would be inefficient with respect to operational costs. Without a focus on the desired state for the organization's analytics functionality, an isolated, prototype-based development model can result in a distributed collection of loosely connected, marginally supported applications that incur increased maintenance costs. It is for that reason that organizations must keep both growth and integration strategies in place throughout each iteration.

Foundational Components of Successful BDA

Processes

Analytics-driven organizations do more than employ the outcomes in their existing decision-making and operational processes. Rather, the most successful firms are those that allow the analytics to drive decision making and operations with little second guessing by managers and staff. In time, this will likely become far more prevalent as the processes and routines of the organization, as well as its products and services, grow increasingly dependent on BDA.[5]

Many organizations will transition from descriptive to predictive to prescriptive analytics, resulting in processes that may become self-optimizing and autonomous. For instance, manufacturing firms will be able to take advantage of prescriptive analytics that determine optimum inventory levels. Transportation firms can use BDA to identify and optimize loads and routes to save the costs of shipping goods.[6] In nearly every industry, the expansion of these projects in the long term will be a driver of process and technological change.

Consequently, the first thing for an organization to do as part of this long-term plan is to identify the anticipated process changes that would flow from the desired strategic plan. In the near term, this long-range vision

Attribute	Top-Down	Bottom-Up
Costs per application	High	Low
Integration with legacy systems	High	Low
Specificity of solutions	Low	High
Expandability	High	Low
Speed of development	Low-Moderate	High
Typical purposes	Enterprise-wide applications and processes; efficient data usage	Specific applications; small-scale deployments; rapid; specialized data

Table 1 — Top-down vs. bottom-up analytics implementation.

should drive much of the innovation planning for BDA projects as changes in these processes are identified as targets for the organization's future operating states. After this has been established, changes in the other components become easier to estimate. Clearly, there is a high degree of correspondence between the choice of technologies and the processes inherent to them. But the desired processes in turn impact not only the technological changes, but also the corresponding changes in staff capabilities, data management, and more.

Although the amount of available data has increased exponentially, the utility of this data has not increased by the same proportion.

Data Management

Once the processes and corresponding technologies are chosen, attention then turns to the data that feeds the BDA processes. For many years, organizations collected mostly structured data. Even the ETL processes employed for enterprise data warehouses were designed to transform much of the data to a more structured format to enable ease of analysis. In recent years, data such as email text, social media posts, and freeform comments have taken the focus as firms attempt to capitalize on the rich insights contained therein. But there are still many challenges to overcome in this area.

For instance, one expert has argued that of the terabytes of data available to a given firm, only a small fraction is actually coded and usable today for the development of algorithms.[7] As a result, although the amount of available data has increased exponentially, the utility of this data has not increased by the same proportion. Other data concerns include the abundance of unstructured data, the inability to match data across tools and applications, and the lack of interoperability between organizations. These challenges will ultimately lead to a suboptimal environment upon which to develop and implement trustworthy insights.

Another ongoing issue is that the data is being housed in multiple systems, making integration difficult. Many times, there is no interoperability between these systems, making comprehensive analytics nearly impossible. For example, many organizations discover that interoperability and integration become a significant hurdle as they participate in mergers and acquisitions. Due to the sensitive nature of some data, many organizations are unable to understand the complexity of the different data types, collection options, and levels of system integration until after the quiet period is done. In these situations, organizations may attempt to determine the extent of the forthcoming challenge through the use of third parties who view both sides of the agreement. Until that point, analytics and data integration are largely infeasible.

Staffing

One of the biggest issues to consider in BDA adoption is the staffing capabilities required to operate a BDA platform. As these new technologies are introduced into the organization, an additional level of knowledge, skills, and expertise is needed. Thus, organizations must evaluate the necessary changes in staffing levels with each iteration. This includes a detailed look at the existing staff to determine whether the organization has the right knowledge, skills, and abilities to take advantage of current and future technologies. The focus is not strictly on the number of people available within the organization and the personnel cost (although these certainly must be taken into account). Rather, the challenge is maintaining the appropriate mix of experience and expertise to manage the technologies required for existing operational demands as well as impending BDA projects.

There are three primary ways to address the shortfall. First, the organization can hire new technical staff. Second, it can retain contract or temporary workers to deal with any short-term gaps in knowledge requirements. Third, the organization can train existing staff to handle the new technology, if there is adequate time. This last approach (talent development) may prove to be the most fruitful in many cases, particularly when we consider Brooks' Law.[8] In many organizations, the temptation is to hire a bunch of staff to handle newer technologies, forgetting that the current employees have a wealth of institutional memory with regard to the specifications, interfaces, workarounds, and even political considerations associated with the installed technologies. Adding more explicit technical knowledge along with this tacit knowledge often proves to be more useful.

Regardless of the solution, an organization must commit to making changes in the knowledge and expertise available at each iteration. Inadequacies lead to inefficiencies or, at worst, an inability to use any new capabilities to their full capacity.

A Case Example

Several of the issues we've been discussing are exemplified by the experiences of an analytics manager we've worked with in the financial services industry. While content with the state of his department's current BDA platform, he expressed a need to stay abreast of the firm's future concerns in order to be prepared to apply newer technologies to respond more effectively to emerging competitive demands. For instance, he spoke of establishing benchmarks for performance in their marketing analytics, including such areas as anticipating customer needs and minimizing physical visits to their facilities. Over time, the standards for performance from their marketing efforts will become more demanding as competitors improve their ability to identify and convert leads into customers. It is likely that meeting these standards will require BDA technologies, processes, and skills that may not be available to the firm at the present time.

Accommodating the new applications within the firm will require several aspects to be dealt with directly. As new applications are integrated into the existing platform, the IT and analytics staffs will have to address changes in the firm's overall data structure, including resolving access to legacy data and validating new data as it becomes available. The new platform also highlights the need to plan for training the current staff on newer technologies (Hadoop, Spark, etc.) while maintaining the skills necessary to manage their current SQL-based environment. While there are several skunkworks projects under development to determine whether the resulting applications will eventually be of use to the firm, the primary focus is on supporting organizational decisions using the existing data warehouse and analytics tools. However, if these technologies were to require separate work groups, or centralized reporting structures, the manager had become aware that it may be necessary to hire additional technical staff over time.

In all, the focus on future strategic changes led to the development of plans to improve the processes, data management, and staffing of the department. By focusing on the issues in this way, the firm is able to manage their current environment while allowing for a gradual transition to their desired future state.

Rising to the Challenges

Clearly, many organizations have realized significant returns using BDA technologies. For instance, the Kroger grocery store chain has been able to increase its customers' direct mail coupon redemption rate to more than 71% (vs. the industry average of 3.7%) by applying analytics to learn as much as possible about each individual customer's likes and dislikes.[9] Similar returns can be found in human resources, auto rental, manufacturing, banking, and many other domains. And yet, successful BDA implementations are difficult, as firms struggle to adapt to the inexorable march of new technologies. With each new technological advance, a new set of challenges arises in each of the three foundational components discussed above: strategic/operational processes, data management practices, and technology staffing. But in time, these get resolved — just in time for new ones to arise.

Successful organizations are able to consistently adapt the components to overcome any challenges. We argue that the following planning process is both ongoing and essential to successful BDA projects:

- Understand where the organization's business models and competitive strategies are headed.
- Determine how BDA can support or lead the organization's strategic and operational evolution.
- Develop a plan for how each component can evolve as the organization's analytics capabilities evolve, specifically:
  - Adapting existing strategic and organizational processes, or building new ones.
  - Sourcing, adapting, loading, and otherwise managing the various data to be used.
  - Comparing current and anticipated levels of expertise and obtaining any new skills.
- Stay vigilant in monitoring new BDA technologies and how they can be integrated into the existing platform (or replace it). The new technologies may also require the adaptation of existing components.
  - In cases where speed to market or speed of development matters, set up proof-of-concept projects to assess the viability of these new technologies and the results they generate.
  - In other cases, the new technologies should be carefully integrated with the existing platform and data structure.
- Ensure that the costs of running each new project do not become excessive due to specialized knowledge, unique data storage, and inadequate maintenance coverage.

- Help users adapt to any changes in decision-making authority, social systems and interorganizational relationships, and any other strategic or cultural factors.

These steps will certainly demand a high degree of attention and ongoing monitoring by the IT staff, senior management, and the analytics team. However, organizations that can simultaneously manage their processes, data, and human resources can expect to not only resolve the current issues, but be better prepared for long-term success from their BDA efforts.

Endnotes

[1]Wynn, Donald, Jr., and Renée M.E. Pratt. "The Promises and Challenges of Innovating Through Big Data and Analytics in Healthcare." *Cutter IT Journal*, Vol. 27, No. 4, 2014.

[2]"Healthcare Analytics Adoption Model." HealthCatalyst (www.healthcatalyst.com/healthcare-analytics-adoption-model/).

[3]Marr, Bernard. "How Big Data Is Changing Insurance Forever." *Forbes*, 16 December 2015 (www.forbes.com/sites/bernard-marr/2015/12/16/how-big-data-is-changing-the-insurance-industry-forever).

[4]McCann, Erin. "A Beginner's Guide to Data Analytics." *Healthcare IT News*, 2 June 2015 (www.healthcareitnews.com/news/beginners-guide-data-analytics).

[5]Davenport, Thomas H. "Analytics 3.0." *Harvard Business Review*, December 2013 (https://hbr.org/2013/12/analytics-30).

[6]Van Rijmenam, Mark. "The Future of Big Data? Three Use Cases of Prescriptive Analytics." Datafloq, 29 December 2015 (https://datafloq.com/read/future-big-data-use-cases-prescriptive-analytics/668).

[7]Sittig, Dean. Interview by the authors. University of Texas, 15 April 2016.

[8]Brooks' Law states that "adding manpower to a late software project makes it later."

[9]Morgan, Lisa. "Big Data: 6 Real-Life Business Cases." *InformationWeek*, 27 May 2015 (www.informationweek.com/software/enterprise-applications/big-data-6-real-life-business-cases/d/d-id/1320590).

Donald E. Wynn, Jr., is an Associate Professor in the School of Business Administration at the University of Dayton. He holds a PhD in business administration from the University of Georgia. His research appears in journals such as MIS Quarterly, MIS Quarterly Executive, Information Systems Journal, Cutter IT Journal, Journal of Organizational and End User Computing, Communications of the AIS, *and* Journal of the Academy of Marketing Science. *Dr. Wynn has published research articles in a number of areas, including open source software, big data and analytics, electronic health records software, technological ecosystems, and research methodologies. He can be reached at dwynn1@udayton.edu.*

Renée M.E. Pratt is an Assistant Professor at the University of Massachusetts Amherst in the Isenberg School of Management in the Operations and Information Management Department. She received her doctorate from Florida State University in management information systems. Dr. Pratt completed research in 2013 on the implementation and post-adoption processes of electronic medical record software under a Fulbright Scholar Grant in Germany. Her research interests include post-adoption diffusion behaviors, traditional and clinical enterprise systems, healthcare IT, and psychological contract. Dr. Pratt's work has appeared in numerous publications, including MIS Quarterly Executive, Cutter IT Journal, Information Systems Journal, Journal of Information Technology, *and* Information Systems Journal of Education, *among others. She can be reached at rpratt@isenberg.umass.edu.*

Enabling Agronomy Data and Analytical Modeling: A Journey

by Mohan Babu K

Agriculture is among the oldest vocations known to mankind. Traditionally, a farmer's decision making is grounded in human knowledge and intelligence that comes from experience (analysis of historical data), intuition (predictive modeling), and insights from such analysis (visualization of such data with recipes and formulas). Agriculturalists are increasingly moving away from depending on empirical knowledge toward working with modern tools and techniques grounded in data and analytical modeling. Such tools and techniques are modernizing the decision making and enabling increasing yields and revenue for farmers.

Farmers and the agricultural companies that service their needs deal with vast amounts of structured and unstructured data. Analysis of such data gathered from across a variety of growers and growing conditions, combined with data from other sources — including satellite and drone imaging, field-level sensors, weather, and other historic data — can provide insights to enable farmers to make timely decisions that can improve their yields and minimize losses due to unpredictable changes in weather.

In this article, I will examine the fast-changing landscape of agronomic data gathering and modeling. I will also look at the enablers for the following kinds of analytics, which require different data inputs and quality of data:

- Investigative (discovery)
- Descriptive (aggregation)
- Predictive (outcomes)
- Prescriptive (available options)

I will also evaluate data integration capabilities required to deliver these analytical capabilities. Gathering such data for decision making is not a trivial challenge that can be addressed with a single solution; rather it is a journey that aims to provide farmers with tools for decision making. I will also highlight some observations and learnings for those striving to enable data for analytics and modeling in agriculture.

Data and Technology Aiding Agriculture

Technology is pervasive in almost all aspects of modern agriculture — from the time a farmer plans the crops for the season, and even after the crops are harvested and leave the farm, through marketing and distribution. At the time of planning crops for the growing season, the farmer takes into account key agronomy inputs, including the long-term weather forecast (when do I plant my crops?), grain price forecast from futures markets (what crops do I plant?), availability of new agriculture technology, including quality seeds and agro-chemicals such as pesticides, herbicides, and insecticides (what technologies do I use to maximize my yields?). To answer some of these questions, the farmer also has to consider other basic inputs like the available land acreage, access to irrigation, and the labor and resources at his or her disposal.

Farmers realize they have enormous amounts of data at their disposal, but they also recognize that analysis of data is not their core competence.

Farmers realize they have enormous amounts of data at their disposal, but they also recognize that analysis of data is not their core competence. They need tools, technologies, and advice to interpret the data that can enable them to make timely decisions. Figure 1 highlights some typical questions for which farmers need answers in order to enable planning during a growing season.

During the growing season, farmers have to continually monitor their fields. The popular "Farm Forward" video from John Deere[1] takes a futuristic view of technologies by integrating information "just in time" for decision making. Many of the technologies highlighted are already being adopted in farming, although end-to-end integration of individual solution components and data sources remains a challenge.

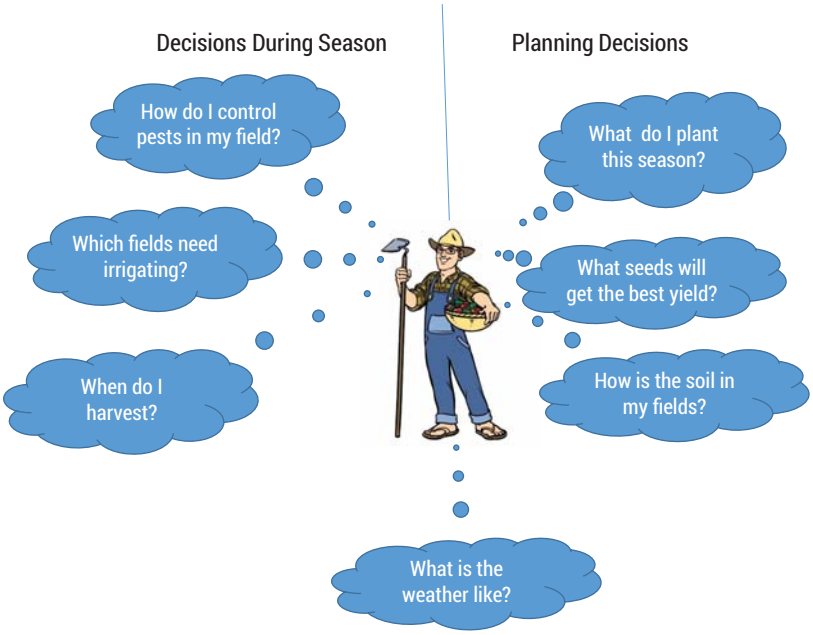


Figure 1 — A sampling of the questions that drive decisions.

In addition to physical monitoring and surveying, farmers are aided by sensing technologies and data feeds from satellites and drones. Structured and unstructured big data from such an Internet of Things (IoT) at farms is increasingly being used to answer questions that may not be intuitive for individuals. For example, the recent case study "Connected cows help farms keep up with the herd" highlights an innovative case:

> SCR Dairy now has about 4 million tags connected to cows around the world, monitoring their activity and wellbeing 24 hours a day. The data generated from the tags is transferred to management solutions that help farmers make better decisions, as well as providing alerts.
>
> "We have alerted farmers of cows having, for example, a prolonged calving, or a difficult labor, in the middle of the night," says Matteo Ratti, vice president of SCR's Cow Intelligence business. "They were able to go out and save the cow. With this technology, farmers get the information they need to manage the herd more efficiently."[2]

During the growing season, farmers also need to continually respond to changes in weather, rainfall, and increase of pests, weeds, and other factors that can impact the growth and yield. On a large farm, the farmer might have to take action on a field level by increasing or reducing irrigation and managing the application of pesticides, herbicides, and insecticides in a controlled manner.

Decision Support Systems in Agronomy

Decision support systems for agronomy take a few fundamental factors as inputs, some of which are within the control of farmers and many which are outside their control. Figure 2 highlights some of the key decisions.

Providing field-specific advice to growers requires aggregating, analyzing, and tailoring agronomy data based on inputs that include local soil, crop variety, weather and environmental conditions, pests, and other inputs. Technology firms across a broad spectrum of the agriculture industry are attempting to aggregate such data from public and corporate sources. A recent *Farm Industry News* article[3] highlights major American companies with solutions for real-time farm management and agronomy.

The goal of most agronomy solutions is similar: they aim to provide field-specific advice, tailored to local soil, crop variety, weather, pests, and environmental conditions for farmers. However, in order to provide such tailored advice, field-by-field data needs to be gathered from the growers and captured in a model, designed in the agronomy solution. In addition, data from external sources, including weather data and forecasts and other agronomic inputs from public sources, needs to be gathered. The end result is a recommendation on potential yields and planning for the following year.

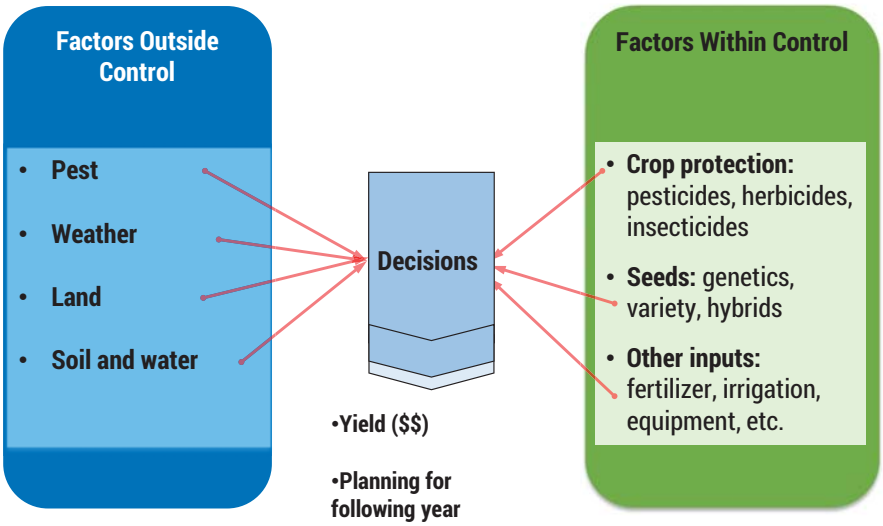


Figure 2 — Inputs for decision support systems.

Opportunities in Data and Analytics

Enabling field-specific advice localized to individual farms and farmers requires the design of processes and systems that can take inputs from structured and unstructured data sources. Such data, when aggregated and analyzed, can provide timely insights and recommendations.

There is a confluence of forces impacting the way consumers interact with information technology, including what some in the industry collectively call SMAC — social, mobile, analytics, and cloud.[4] SMAC is also being seen as an opportunity in agribusiness as companies, government organizations, and others interact with and influence farmers and growers. Agricultural research institutions, government agencies (like the US Department of Agriculture), and agribusiness companies actively engage with farmers using social media accounts on Facebook, Twitter, LinkedIn, and so on. Many of these accounts are supported by people and infrastructure to provide timely crop and region-specific information on weather, pests, and other factors that farmers watch closely. These social media accounts become really active during the growing season.

Farmers, especially in Western economies, have been early adopters of social media as they engage and interact with others via tweets, knowledge-sharing blogs, Facebook groups, and other online forums using their mobile devices. Farms operate outside of urban hubs and may not have the best network and broadband access, but this has not constrained the adoption of SMAC tools and technologies. Rural communities, especially in developed economies, are increasingly deploying their own Internet network connections and hubs.[5] In developing economies like India, SMAC also includes the use of more pervasive technologies like SIM cards on low-end cellphones. A recent *Ars Technica* article highlights how SMS technology is being used: "the device leverages the core functions of a SIM card (transmitting calls and texts) to deliver free voice messages to farmers, offering updates relating to growing and selling crops."[6]

Modern agriculture has long relied on weather and agronomy data gathered from satellites, weather stations, and other publicly managed platforms. Such information, gathered from government, research institutions, and other public sources, has been made available to growers freely. Design of analytics and agronomy decision support systems require deeper understanding of agriculture data available from public, corporate, and farm data sources:

- **Corporate agronomy data.** Information on sales and marketing, pests, and agronomic inputs are constantly analyzed by organizations involved in the production and supply of seeds and crop protection chemicals.
- **Farm data.** Individual farms and farmers generate a lot of data from their operations, including data on their land, soil, seeds, chemical, fertilizer, water, and other inputs used during the growing season; agronomic protocols they have adopted; and details of their crop and historic yield, along with local growing conditions.
- **Public data.** Public data (including data from national, state, and local government agencies), weather and agronomy data, market analysis and forecasts, and other information may either be available freely or sourced from data aggregators.

- **Emerging data.** Adoption of consumer-centric uses of the IoT continue to await mass adoption, but farmers are already beginning to leverage data from farm-based sensors and drones.[7]

Integrating data from across traditional and emerging data sources requires an understanding of data formats, types, frequency, aggregation, and translation of such data. Rules and regulations with respect to data ownership and stewardship — especially of farm-specific data — vary across countries, states, and provinces. Therefore, agronomy data scientists also need to be cognizant of regulatory requirements that guide the storage, aggregation, and retrieval of such data.

A Framework for Analytics in Agronomy

Innovative data aggregators, organizations, and scientists are applying different types of analytic techniques such as investigative data discovery, descriptive data aggregation, predictive analytics focused on outcomes, and other prescriptive techniques. Figure 3 shows a framework for analytics in agronomy. The framework is designed to enable architecturally significant use cases, including:

- **Reporting.** Farming operations require daily/periodic reports on a number of topics, including weather, grower and subcontractor results, and information required to follow up on planning versus actual activities. Such reports support regular farm operations and help with planning of future activities. Farms also have to maintain reports and data on seeds and applications of pesticides, herbicides, insecticides, and other treatments. Such reports are needed for cost and yield analysis and may also be required for inspection by federal and state farming regulatory authorities.
- **Dashboards.** The other major reporting capability is to enable dashboards for visualization and analysis. This includes dashboards for planning activities like crop planting and diagnostics of factors that could impact the quality and yield performance. Data and images gathered from satellites, drones, and sensors can also be visualized against field-level coordinates to observe the progress of crop growth and plan any required course corrections. Farmers may also require the ability to extract and transmit such data to agronomists and other advisors.
- **Discovery.** Support for predictive analytics is another major capability being designed into the agronomy framework to enable diagnostics and search and data exploration. Such predictive analytics require historical data to observe variance between recommended and actual yields and other limiting factors. For example, analysis of the data may highlight a farm plot that consistently yields better results than others in the vicinity that don't get similar inputs. The farmer and agronomists can then drill down and review other factors regarding the plot to understand this positive variance and whether it can be replicated across the farm.

Dashboards enabled by predictive analytics are already starting to pay dividends in farming operations. For instance, a recent Reuters article quotes farmer Juergen Schwarzensteiner, who rotates corn, potatoes, and grains at a 970-hectare farm in Bavaria using satellite maps and analytics software: "This plot has had top yields consistently over the years, [and] I used to just say, that's great.... Then we got the digital maps, and differences became apparent that were not clear to the eye before."[8] Using digital dashboards, farmers like Schwarzensteiner are able to view color-patterned digital maps that highlight discrepancies in plants growing in plots across fields even half a mile apart. Such dashboards "aim to provide farmers with individualized prescriptions on how to work each field down to a fraction of an acre, using data they have collected on soil and weather conditions."[9]

At the core of the framework are structured and unstructured data sources. Agribusiness organizations, government agencies, and other research organizations generate reports and transactional data in formats that can be stored and retrieved from relational, structured databases. Such transactional and reference data may exist in databases within software applications running commercially developed databases like IBM's DB2, Microsoft' SQL Servers, or Oracle. Such data can be cataloged, indexed, and queried using well-understood tools and techniques.

Social media, satellites, drones, and sensors also generate vast amounts of unstructured and big data that may include images, text, and other data structures. Emerging big data analytic techniques are being applied to make sense of this data. Traditionally, farmers have applied new techniques — such as new seeds, pesticides, herbicides, and so forth — to a small plot to observe optimal yields. Instead of such empirical analysis, which takes time, farmers are also embracing results from analysis of large, real-world data sets from public sources. Analysis of such big data can produce reliable recommendations much more quickly.

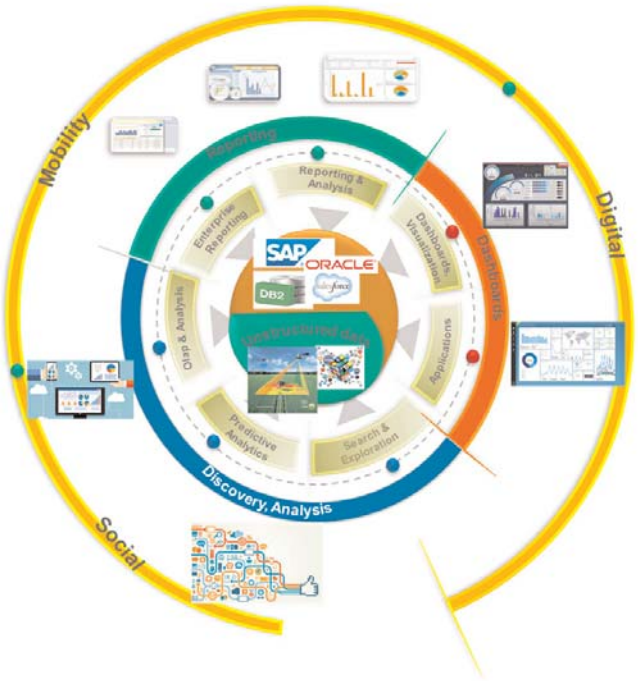


Figure 3 — A framework for analytics in agronomy.

Case In Point: Crop Risk Management Enabled by Analytics

Agribusiness companies thrive by identifying optimal agronomic conditions to provide better yields for their customers, the farmers. Such advice had been based on the empirical knowledge of individual agronomists and researchers. Now these companies have begun leveraging insights based on analytics of data from field trials, crop properties, and other agricultural inputs to identify solutions that can provide consistent and optimal yields. Identifying and leveraging such insights to advise farmers also makes good business sense: growers who enjoy consistently high yields are willing to pay a premium for crop protection chemicals and high-quality seeds.

HYVIDO® barley is one of the first solutions from Syngenta to be backed by a Cashback Yield Guarantee.[1] Based on extensive reviews of field trial data, weather, protocols, and other grower inputs, it was determined that optimal use of HYVIDO barley solution can guarantee a yield increase for growers. Thus, the company is able to offer a guarantee backed by data and analytics. The proposal offers a cashback guarantee, about £60/hectare (approximately US $32/acre), to ensure a half-ton yield boost from growing hybrid winter barley. Growers who sign up for the program get a £60/hectare refund on Syngenta's HYVIDO hybrid barley varieties if certain agronomic conditions are fulfilled.[2]

Other agribusiness companies like Monsanto have also rolled out risk management and yield assurance tools. For instance, Monsanto's Biotech Yield Assurance Program for Farmers, "links an insurance policy to the yield benefits of Genuity® Roundup Ready 2 Yield® soybeans, Genuity SmartStax,™ and SmartStax™ corn."[3]

The genesis of such "crop yield guarantee" solutions lies in extensive analysis of digital agronomy data, including historical yields, data from reference farms, and other inputs. Based on such analysis, agribusiness firms may determine that use of the prescribed protocol can lead to a better yield, in turn generating higher returns for growers. The goal is to try to replicate such success for other crops in other regions around the globe.

[1] Impey, Louise. "Cash Back If Hybrid Barley Doesn't Come Up Trumps." *Farmers Weekly*, 17 February 2013 (www.fwi.co.uk/arable/cash-back-if-hybrid-barley-doesn-t-come-up-trumps.htm).

[2] Impey (see 1).

[3] "Monsanto, ARMtech Unveil Biotech Yield Assurance℠ Program for Farmers." Press release, Monsanto, 10 June 2010 (http://news.monsanto.com/press-release/monsanto-armtech-unveil-biotech-yield-assurancesm-program-farmers).

An article from the International Center for Tropical Agriculture illustrates how scientists at the center have applied big data tools to pinpoint strategies that work for small-scale farmers in a changing climate:

> Big Data on Colombian rice comes from commercial fields in ambient weather, a stark departure from other research.... With good analytics, this produces nuanced, reliable recommendations much more quickly. It also gives rice breeders the feedback they need to develop climate-adapted lines.[10]

The challenge is to aggregate data from disparate sources in different formats to draw inferences.

Data aggregated from different sources needs to be analyzed, visualized, and used for reporting. Such data may be designed to persist in reporting tools temporarily or in a fit-for-purpose data warehouse designed for agronomic reporting. For instance, a large farming operation, with hundreds of acres of land spread over a large geographic area may benefit from analysis of aggregated data, especially if trends from one field can predict outcomes in other fields. The aggregated data also needs to be localized. Rick Murdock, head of Ag Connections, a wholly owned subsidiary of Syngenta, explains: "We believe spatial agronomic recommendations are local and need to be driven by local retail agronomists, consultants, or grower agronomists: We know crops grow best when they are seen by the agronomist!"[11]

Many of the techniques and solutions highlighted in the framework, including tools for data analysis, reporting, visualization, and aggregating big data with data across disparate sources, are already available. These are being used in other industries and to solve individual agronomy problems. The challenge is to aggregate data from disparate sources in different formats to draw inferences. Among the difficulties in aggregating data from different sources is the need to clean the data. A few key assumptions on data cleansing include:

- **Data is generally not cleansed at the source or during the data collection stage.** It is assumed that such data from different organizations and source systems is formatted and referenced according to its individual requirements. The data may not be designed with a common taxonomy, and even the metadata and units of measure may be different.
- **Data cleansing could be done using automated methods, but this requires some manual effort and standardization of business rules.** Data analysts need to understand the sources of the data, data definitions, and metadata, and based on such understanding, they can plan to translate and cleanse the data after retrieval.
- **Cleansed data can be stored as published data for various visualization or analytical purposes.** Such cleansed data may have to be reviewed periodically, as it might go stale.

Conclusion

In this article, I have discussed some of the traditional and emerging data analysis and analytic techniques being applied to enhance decision making in farming. Farms operate in a variety of topographies, weather conditions, and geographies across the globe. Farms also operate at a variety of scales, ranging from small subsistence farms to mega farms spanning thousands of hectares. Given the variety of conditions where farms operate, frameworks for tools and technologies to support analytics in agronomy need to be customized to specific requirements, which is an ongoing journey.

I also introduced a framework for generating localized advice and information for individual farms based on data gathered from public and corporate data sources. Current advancements and digitization in agronomy are geared toward increasing yields at large farms that specialize in a few key field crops. Such farms also have the means at their disposal to invest in tools, technologies, and data gathering. I believe that such an agronomy framework will eventually scale down to benefit small-scale and subsistence farming practiced in developing economies as well.

Acknowledgment

I would like to acknowledge my colleagues Stephen Smith and Amar Singh for ideation and research for this article.

Endnotes

[1]John Deere. "Farm Forward" (video). YouTube, 17 July 2013 (https://www.youtube.com/watch?v=t08nOEkrX-I).

[2]Heikell, Lorence. "Connected Cows Help Farms Keep Up With the Herd." *Microsoft News*, 17 August 2015 (https://news.microsoft.com/features/connected-cows-help-farms-keep-up-with-the-herd/#sm.00019ajpzz1dgzdqrrxkm3wtxff3j).

[3]Vogt, Willie. "Putting Data to Work — Part 2." *Farm Industry News*, 27 April 2016 (http://farmindustrynews.com/precision-farming/putting-data-work-part-2).

[4]Evans, Nicholas D. "SMAC and the Evolution of IT." *Computerworld*, 9 December 2013 (www.computerworld.com/article/2475696/it-transformation/smac-and-the-evolution-of-it.html).

[5]Turk, Victoria. "This Rural Community Is Building Its Own Gigabit Internet Network." Motherboard, 7 May 2014 (http://motherboard.vice.com/read/this-rural-community-is-building-its-own-gigabit-fibre-network).

[6]Tveten, Julianne. "How a Simple SIM Card Makes Farmers More Efficient — And Possibly Saves Lives." *Ars Technica*, 20 March 2016 (http://arstechnica.com/gadgets/2016/03/how-a-simple-sim-card-helps-farmers-navigate-changing-climates-and-markets/).

[7]Wihbey, John. "Agricultural Drones May Change the Way We Farm." *The Boston Globe*, 23 August 2015 (www.boston-globe.com/ideas/2015/08/22/agricultural-drones-change-way-farm/WTpOWMV9j4C7kchvbmPr4J/story.html).

[8]Burger, Ludwig. "Digital Farming Could Spell Shake-Up for Crop Chemicals Sector." Reuters, 4 May 2016 (www.reuters.com/article/us-farming-digital-idUSKCN0XV0KP).

[9]Burger (see 8).

[10]"Big Data for Climate-Smart Agriculture." International Center for Tropical Agriculture, Research Program on Climate Change, Agriculture and Food Security (https://ccafs.cgiar.org/bigdata#.V049B_krLIU).

[11]Vogt (see 3).

Mohan Babu K is an Enterprise Architect (IS) at Syngenta, a multinational agribusiness company based in Greensboro, North Carolina. He has spent two decades in technology management and has gained a strong insight into the lifecycle of portfolio management and the global delivery model. Having lived and worked in a dozen countries on three continents, he has also gained an international perspective on business and society.

Mr. Babu K's viewpoints and papers have been published in several international technical and nontechnical journals, including Cutter IT Journal, Business Integration Journal, Research-Technology Management, IEEE Computer, Computerworld, ACM Ubiquity, *and* Sourcingmag, *among others. He is the author of a book on globalization titled* Offshoring IT Services: A Framework for Managing Outsourced Projects. *He can be reached at mohan@garamchai.com, LinkedIn: www.linkedin.com/in/mohanbabuk/.*

Cutter
IT Journal

About Cutter Consortium

Cutter Consortium is a truly unique IT advisory firm, comprising a group of more than 100 internationally recognized experts who have come together to offer content, consulting, and training to our clients. These experts are committed to delivering top-level, critical, and objective advice. They have done, and are doing, groundbreaking work in organizations worldwide, helping companies deal with issues in the core areas of software development and Agile project management, enterprise architecture, business technology trends and strategies, enterprise risk management, metrics, and sourcing.

Cutter offers a different value proposition than other IT research firms: We give you Access to the Experts. You get practitioners' points of view, derived from hands-on experience with the same critical issues you are facing, not the perspective of a desk-bound analyst who can only make predictions and observations on what's happening in the marketplace. With Cutter Consortium, you get the best practices and lessons learned from the world's leading experts, experts who are implementing these techniques at companies like yours right now.

Cutter's clients are able to tap into its expertise in a variety of formats, including content via online advisory services and journals, mentoring, workshops, training, and consulting. And by customizing our information products and training/consulting services, you get the solutions you need, while staying within your budget.

Cutter Consortium's philosophy is that there is no single right solution for all enterprises, or all departments within one enterprise, or even all projects within a department. Cutter believes that the complexity of the business technology issues confronting corporations today demands multiple detailed perspectives from which a company can view its opportunities and risks in order to make the right strategic and tactical decisions. The simplistic pronouncements other analyst firms make do not take into account the unique situation of each organization. This is another reason to present the several sides to each issue: to enable clients to determine the course of action that best fits their unique situation.

For more information, contact Cutter Consortium at +1 781 648 8700 or sales@cutter.com.

The Cutter Business Technology Council

The Cutter Business Technology Council was established by Cutter Consortium to help spot emerging trends in IT, digital technology, and the marketplace. Its members are IT specialists whose ideas have become important building blocks of today's wide-band, digitally connected, global economy. This brain trust includes:

- Rob Austin
- Ron Blitstein
- Tom DeMarco
- Lynne Ellyn
- Vince Kellen
- Tim Lister
- Lou Mazzucchelli
- Robert D. Scott