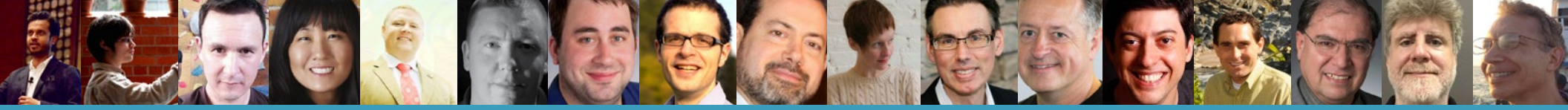
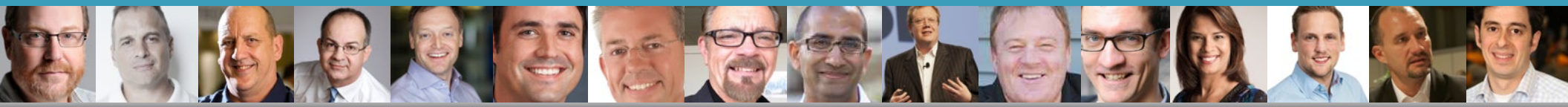


Data *Disruption*

Technology Platforms



Mighty Guides

INTRODUCTION

It is impossible to ignore the fact that we live in a data-driven civilization. Not only is the amount of data in the world doubling every two years, but the percentage of these data that are becoming valuable because of advanced analytics is also growing. The entire field of data capture and analysis is evolving so rapidly that organizations have difficulty keeping up. Yet data-driven business processes, competition, and the rewards of faster and more intelligent operations leave us with no other choice.

For a long time, our ability to capture data outpaced our ability to process it. This meant that large quantities of data were stored in data warehouses until some future time when tools would be available to find value in them or until they were discarded all together. Several things have happened in recent years to change this dynamic. One is the exponential growth in data; the other is the emergence of new platforms and technologies that make it possible to process data sets of almost unlimited size economically while lowering the cost and increasing the speed of analysis. These elements, combined with new analytic techniques and a growing use of machine learning to accelerate analytic methods, is changing almost every aspect of our lives.

To gain a fuller understanding of how modern analytical methods are being used in visible and not-so-visible ways, we approached data analytics experts from many fields and industries. I asked them to contribute essays about their experiences applying big data analytics. This e-book is a compilation of those essays. In it you will find discussions about new analytics technologies, how organizations can more effectively use their data assets, and many interesting use cases. The essays have been grouped into five sections:

- **Business Change.** Essays in this section speak to how advanced analytics are changing the way businesses operate. It is much more than a story about increased productivity and efficiency: it is a story about the complete transformation of traditional business models into something new and totally data driven.
- **Technology Platforms.** Essays in this section take a closer look at some of the tools and platforms that are making advanced analytics economical for organizations of all sizes.

- **Industry Examples.** This section continues the discussion of transformative analytics technologies in the context of specific business and public-sector use cases.
- **Research.** This section focuses on how new-age analytics are changing the way scientists are conducting research and how they are speeding knowledge acquisition.
- **Marketing.** This section focuses on advanced, analytics-driven marketing strategies and techniques. These techniques are being used for everything from brand marketing to personalization to public relations to attribution techniques that enable companies to analyze their most effective marketing activities in real time.

It is my hope that assembling knowledgeable insights and experiences from so many different perspectives will provide a valuable glimpse into this rapidly evolving technology. I have found many of these essays both eye-opening and thought provoking. There is no question that advanced analytics will continue to play an increasingly important role in business, government, health care, knowledge acquisition, and a broad spectrum of human endeavor.

All the best,
David Rogelberg
Publisher

Mighty Guides

Mighty Guides make you stronger.

These authoritative and diverse guides provide a full view of a topic. They help you explore, compare, and contrast a variety of viewpoints so that you can determine what will work best for you. Reading a Mighty Guide is kind of like having your own team of experts. Each heartfelt and sincere piece of advice in this guide sits right next to the contributor's name, biography, and links so that you can learn more about their work. This background information gives you the proper context for each expert's independent perspective.

Credible advice from top experts helps you make strong decisions. Strong decisions make you mighty.

Technology Platforms

Scott Shaw
Hortonworks.....5

Justin Langseth
Zoomdata.....7

Sean Owen
Cloudera.....9

Mark Himelstein
Graphite Systems, Inc.....11

Adam Gibson
Skymind.....13

George Gilbert
Wikibon.....15

Abhishek Mehta
Tresata.....17

HIGHER PROCESSING SPEED AT LOWER COST IS CHANGING EVERYTHING

SCOTT SHAW

Solutions Engineer, Hortonworks

Scott Shaw has more than 15 years of experience in data management and business analytics. He has worked as an Oracle and Microsoft SQL Server database administrator; he was a principal consultant and headed a data management practice. Scott worked closely with clients to design and model complex data warehouses. He now works as a solutions engineer for Hortonworks, providing architecture and deep learning around open source Apache Hadoop. Scott lives in Saint Louis with his wife and two children.

Twitter | Website

Download the full e-book: Data Disruption

Companies have been processing, dissecting, aggregating, and analyzing data for decades. Massive database systems or mainframes take in large data sets, and business analysts use various tools to drill through the data to create operational or forecasting reports. These reports drive the business and help companies determine the current state of affairs and develop insights into new business opportunities. Database systems excel at putting data sets into memory for fast processing. As memory prices drop, larger and larger data sets are pushed into growing memory caches. None of this is especially new, and it is still the primary means most companies use for their analytics. So what has changed?

KEY LESSONS

1. HADOOP PROVIDES INEXPENSIVE, HORIZONTAL, AND LINEAR SCALABILITY AT THE STORAGE AND COMPUTE LAYERS.
2. DIFFERENT DATA SOURCES CAN NOW BE COMBINED AND VIEWED THROUGH A SINGLE PANE OF GLASS.

We have undergone a paradigm shift in data platforms. This shift started at the lowest levels: storage and compute. Apache Hadoop provides inexpensive, horizontal, and linear scalability at the storage and compute layers, which means that companies now have an analytics platform that easily stores and processes data in parallel across all systems. Complex data-management systems (typed data types, relational constraints) no longer hinder data ingest, and tool set and scalability limitations no longer hinder data egress.

> “Hadoop allows for a mixture of data access (batch, interactive, real time) all on the same platform, all acting on the same data.”

HIGHER PROCESSING SPEED AT LOWER COST IS CHANGING EVERYTHING

Traditional relational systems were built specifically for random read–write, row-level access. If you wanted to provide analytics that required longer, sequential access, you had to build a new system. Hadoop allows for a mixture of data access (batch, interactive, real time) all on the same platform, all acting on the same data.

With this capability, web log data can live next to customer relationship management data. Sensor data can live next to internal product data. Streaming data can be stored next to relational data. These data sources can then be combined and viewed through a single pane of glass. In machine learning, these different types are referred as *features*. The idea is that the more features you introduce, the better your model. Take for example home energy use. Thermostat data are fine, but these data do not complete the picture. What if you also introduced heating, ventilation, and air conditioning sensors; weather data; historical energy use; and square footage data? By adding features, you build more accurate models. The challenge is that adding features increases the demand on processing.

Hadoop solves this issue in two ways. First, it provides the means to store massive amounts of data at scale, which is the foundation of the model. Second, through a compute process such as Spark, companies can quickly build and train models in memory. Now, the system can look at all those data in real time and provide more accurate projections for system maintenance, power consumption, and other control functions.

> “Companies can build more accurate predictive models that anticipate action in the marketplace.”

With faster data processing tools, not only can companies react faster to a changing market, but they can begin to run their analytics pretransaction. Companies using Hadoop and in-memory processing can build more accurate predictive models that anticipate action in the marketplace. Companies can promote to customers most likely to buy their products. Companies can target activities most likely to be fraud. Physicians can engage with patients who are most likely at risk. In a sense, the flexibility of distributed storage and the speed of in-memory processing have created a competitive race for companies in which the realization of future gains is determined by how far and how deeply they look into the past.

JUSTIN LANGSETH
CEO,
Zoomdata

Justin Langseth is CEO of Zoomdata, developers of the fastest visual analytics for big data. Zoomdata is Justin's fifth startup: he had previously founded Strategy.com, Claraview, Clarabridge, and Augaroo. Justin is an expert in big data, business intelligence, text analytics, sentiment analytics, and real-time data processing. He graduated from MIT with a degree in management of information technology and holds 14 patents. He is eagerly awaiting the singularity to increase his personal I/O rate, which is currently and frustratingly pegged at 300 baud.

Twitter | Website

Download the full e-book:
Data Disruption

Data these days are so big and fast that, for me, the only practical way to handle them is as a continuous stream, end to end. That, of course, is a major disruption to the way we have been doing things.

One casualty of that disruption is the traditional extract, transform, load method of batching and transforming data. If you depend on what used to be called *batch windows*, where you stop everything and cascade in a bunch of integration steps, you will find that batch windows fail if they are overloaded with rapid-fire massive data. Then, you must re-run and debug the batches.

Data warehousing—the consolidation of data in a single, carefully designed storage system—is also fading away. By harnessing the power of Apache Spark, Apache Hadoop and big data, and then in-memory processing to logically join data as needed, there is no need to gather it and store it in advance.

> "Afterward, the customer enjoyed a newfound capacity to handle, manage, and act on IoT data."

KEY LESSONS

1. TREATING IT AS A CONTINUOUS, REAL-TIME STEAM IS ONLY PRACTICAL WAY TO HANDLE BIG DATA.
2. THE OLD APPROACHES, SUCH AS THE ETL METHOD OF BATCHING AND TRANSFORMING DATA, AND DATA WAREHOUSING ARE FADING AWAY.

A DATA RIVER RUNS THROUGH IT

My company's technology has Spark embedded as a native application. It is built to be completely stream-native, end to end. Because we are a startup and because the open source Spark environment is so robust these days, we were fortunate that we could develop that way. It allows us to help customers deal with real-time sensor, equipment, and device data from the Internet of Things (IoT).

One client is a computer device company that has hundreds of thousands of devices deployed globally. Those devices generate massive continuous streams of data daily. Our customer needed a way to harness and analyze those data, to identify device anomalies and problems and respond in real time. Previously, the client would have spent millions building a system to ingest all those data nightly, and then carefully massage, clean, and organize it to eventually generate actionable information. Such an approach was no option for our client. The data transport and extraction, transformation and enrichment, storage, reporting, and visualization all needed to be continuous and real time.

> "If your system hasn't been revamped in the past two years, it probably needs to be re-architected now."

To facilitate that goal, our client ripped out and replaced almost everything it had in place, implementing newer technologies capable of handling and distributing unprecedented data flows. But afterward, the client enjoyed a newfound capacity to handle, manage and act on its IoT data.

Data volumes and speeds are growing exponentially just as computer-processing capabilities are rapidly expanding. The old approaches just can't compete. With that in mind, my advice is that if your system hasn't been revamped in the past two years, it probably needs to be re-architected now. Things have changed that much that fast.

If you do consider building something new or simply decide to refactor what you already have, think hard about ways to keep your data moving, end to end, as a stream. If you accomplish that goal, get ready for a huge surge of power.

SEAN OWEN
Director, Data Science, Cloudera

Sean Owen is director of Data Science at Cloudera in London. Before Cloudera, he had founded Myrrix Ltd (now, the Oryx open source project) to commercialize large-scale real-time recommender systems on Apache Hadoop. He is an Apache Spark committer and co-wrote *Advanced Analytics on Spark*. Sean was a committer and vice president for Apache Mahout and co-author of *Mahout in Action*. Previously, he was a senior engineer at Google.

Twitter | Website

Download the full e-book:
Data Disruption

For me, Peter Norvig's "The Unreasonable Effectiveness of Data" was the turning point. Norvig is the director of research for Google, and his 2009 essay made a convincing argument that using sophisticated models to analyze small data sets is ineffective compared to applying simple algorithms to increasingly large data sets. Big data, in other words, wins.

That was a rallying cry for a new line of advance. Instead of thinking about how to marginally improve algorithms, why not scale up simple approaches and make them applicable to big data sets distributed over multiple nodes?

That insight got me interested in Apache Hadoop, a Java-based open source software framework for distributed storage and processing, and Apache Mahout, a project for building an environment for creating scalable machine-learning applications quickly. Roughly speaking, those have been the twin themes of my work ever since.

To illustrate what it all means, let's look at recommender engines.

> "Using sophisticated models to analyze small data sets is ineffective compared to applying simple algorithms to increasingly large data sets."

KEY LESSONS

1. **BIG DATA TRUMPS SMALL DATA, EVEN WHEN RUN THROUGH SIMPLE ALGORITHMS.**
2. **KNOW BEFORE YOU BUY WHAT *SUPERFAST* REALLY MEANS FOR YOUR BUSINESS.**

In early iterations, Amazon's recommender engine was based on little more than purchase history: you bought this, so you might be interested in that. But purchase history is only one data point and not a particularly useful one. As a result, the engine's predictions were sometimes wildly off the mark.

Over time, storage and memory became less expensive, so it became easier to aggregate and analyze much more data—every click on an Amazon customer's browser, for example. Many more bits of information were fed into a recommender engine. That engine learned and became much smarter with every click.

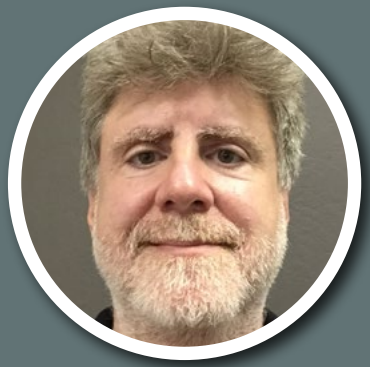
> "Caching data in high-speed flash storage or memory is simply a good thing to do."

This is one example of superfast in-memory computing's disruptive influence. It wouldn't be possible without the increasingly abundant resources that allow us to access and exploit massive data sets. A few years ago, even if I knew how they might benefit me, it didn't matter. Big data sets were just too massive to exploit. Therefore, caching data in high-speed flash storage or memory is simply a good thing to do. It avoids the transfer of files across networks and keeps files off slower-spinning disks, making retrieval much faster. That improvement has qualitatively changed some technology and network design choices.

The human factor is important, too. Clerical grunt-work tasks that highly paid analysts had once performed can be done through brute computational force; tuning model parameters, and selecting model features. Now, we can just empirically let the machines decide. That translates into savings in human time and avoidance of human error. It makes the machine an extension of human talent rather than making analysts slaves to their machines.

If all this sounds exciting, it should. But a word of caution. You will hear a lot of market pitches about real-time business and business transformation that follow in the wake of a superfast, in-memory computing investment. The story there is real—more speed, more resources, and bigger scale can concretely change your business. But think strategically. Know before you buy what *superfast* really means for your business before you take the leap.

THERE IS NO FREE LUNCH



MARK HIMELSTEIN
Chief Performance Architect,
Splice Machine, Inc.

Mark Himelstein is the chief performance architect at Splice Machine, the premier scale out RDBMS. Previously, he was CTO and co-founder of Graphite Systems, Inc. Prior to joining Graphite, Mark was the CTO of Quantum Corp.; before that, he was vice-president of Solaris Engineering at Sun Microsystems, where he led several major revision releases of Solaris and spearheaded the development of the DTRACE, ZFS, and Zones features.

Twitter | Website

Download the full e-book:
Data Disruption

Two years ago, a malware attack on a major US retailer swept up about 40 million debit and credit cards—a massive security lapse that was entirely preventable. What is sad is that the retailer had all the data in its system that it needed to detect and prevent the breach: it simply did not analyze and act on those data.

The business world is moving toward querying structured and unstructured data simultaneously. This retailer's case is just an extreme example: it relied on security software that could not scale with the volume of unstructured events. The company never saw what was coming. Sadly, this is not the only example. One of the world's biggest insurance companies told us that it detects a billion security events a day. How many does it hold onto for later analysis? Zero. It tosses all of them out at the end of each day, for similar reasons.

Many companies like these lack the advanced software and hardware infrastructure they need to analyze and report massive, near–real-time structured and unstructured queries simultaneously. So, here is my point: if you plan to ride the big data wave of industry disruption without having it capsize your business, you must do the hard work. Look beyond the current crop of off-the-shelf, big data analytics solutions for both the hardware and software employed.

> "The retailer had all the data in its system that it needed to detect and prevent the breach: it simply did not analyze and act on those data."

KEY LESSONS

1. THE BUSINESS WORLD IS MOVING TOWARD QUERYING STRUCTURED AND UNSTRUCTURED DATA SIMULTANEOUSLY.
2. HARDWARE-AWARE, HOLISTIC SOLUTIONS ARE THE KEY TO TRANSFORMING BUSINESS.

THERE IS NO FREE LUNCH

One problem with current-generation platforms like Apache Hadoop and Spark—which, incidentally, are excellent for certain jobs—is often they are not quite superfast enough. Hadoop, for instance, is written in Java—a language that does not currently take advantage of the latest high core count extreme NUMA multi-socket computers that represent the current class of enterprise-level processing microarchitecture such as Intel Ivy Bridge and Haswell. Traditional languages like C and C++ can take full advantage of today's advanced processors.

Still, everyone expects that by implementing Hadoop or Apache Spark on enough nodes he or she will handle big data. It rarely works that way. Customers tell me that they are not fulfilling their service level agreements: queries that must be done in less than one minute take an hour. Queries that need to be done in 10 seconds take five minutes. These days, that hurts.

“Everyone expects that by implementing Hadoop or Apache Spark on enough nodes he or she will handle big data. It rarely works that way.”

A hardware-aware, holistic solution is my answer to transforming the way business decisions and therefore money will be made in the future. This is not free or easy. Developing the right solution for your problem takes hard work.

It is no surprise, then, that my company is working on such solutions. But we are not alone. An entire set of companies is working to exploit the flexibility of next-generation software interfaces while respecting and taking advantage of the latest hardware. You would do well to research them all.

People are leaving both security and moneymaking opportunities on the table because their common-wisdom infrastructure choices cannot accommodate big data computing. Many don't even know it. They are trying to do rocket science with only a ruler.

Bottom line: there is no free lunch.

ADAM GIBSON
Co-founder & CTO, Skymind

Adam Gibson is the co-founder of Skymind and creator of the open source libraries Deeplearning4, a distributed deep-learning framework for the Java Virtual Machine (JVM), and ND4J, a scientific computing library for the JVM. He is the author of *Deep Learning: A Practitioner's Approach* (O'Reilly, forthcoming) and currently an advisor to the data science master's program at GalvanizeU. He studied computer science at Michigan Technological University and lives in the Bay Area, near San Francisco.

Twitter | Website | Blog

Download the full e-book:
Data Disruption

Deep learning is a form of machine perception that uses neural networks and huge data sets to teach computers how to solve problems in near–real time through classification and clustering.

OK, but what does that really mean? Say you have a big stack of family photos, a percentage of which have the names attached. A deep-learning framework will rapidly figure out which names go with all those faces. That's *classification*. Now, say you have masses of unlabeled data, like years' worth of raw newspaper articles or a mass collection of unidentified photos. The framework will identify which are the most similar. That's *clustering*.

It is not unlike human perception. Sensory data come in; the brain adds layers of meaning to them, and then decides that the haze of photons in front of me represents my mother. That, too, is an exercise in classifying.

Real-time data processing is the Holy Grail. After all, when we talk about making data go fast, we are really talking about our ability to respond to the world quickly.

> "If you are a bank, it matters a lot whether you identify fraud now or a month from now."

KEY LESSONS

1. **DEEP LEARNING IS A FORM OF MACHINE PERCEPTION NOT UNLIKE HUMAN PERCEPTION.**
2. **DEEP LEARNING IS A SIMPLE, POWERFUL IDEA THAT CAN INCREASE PROFITS, CUT COSTS, AND HELP CREATE NEW PRODUCTS.**

We are talking about disruption. Take a few random examples that we think will benefit hugely from deep learning:

- **Banking.** If you are a bank, it matters a lot whether you identify fraud now or a month from now. A month from now, you will have lost a lot of money.
- **Security.** If you are a security firm or government agency monitoring an airport, it matters a whole lot if you identify a person of interest now or five hours from now—after that person is on a plane over the ocean.
- **E-commerce.** E-retailers want to serve the customer the right ad at the right time, preferably in the instant before a purchase decision. If they only figure out what the customer wanted to see after the customer has left their website, they fail. Getting the customer now is what counts, especially if the customer is using a mobile device with a geolocation app and is walking just outside the store.
- **Automotive.** If you're a carmaker, it matters a lot whether you can identify risks and respond immediately. For example, pedestrians and other vehicles have to be recognized in real time – not minutes later.

> "One of the smartest things you can do is engage the open-source community."

Deep learning can work in various ways. I can best describe our platform. In our case, deep learning basically serves as an analytics layer on top of Apache Spark and Hadoop. Hadoop is a kind of data-management platform, while Spark is a form of MapReduce. Those technologies help orchestrate distributed computing, which is basically synonymous with "fast." I see three ways in which that combination affects business:

- By increasing profits or cutting costs
- By serving as the basis of powerful new products
- By associating the company in the public's eye with innovation—something many businesses feel is important. You have to stay ahead of the curve.

Deep learning is a simple, powerful idea. The trick, if you're going in that direction, is to be smart about it. One of the smartest things you can do is engage the open-source community. Hire people who know how the community operates and who can contribute to deep learning's open-source development. That way, your company can gain credibility with the vitally important open-source community while making sure its needs are met as this powerful platform evolves.

THE FUTURE DEMANDS TECHNOLOGICAL MUSCLE

GEORGE GILBERT

Senior Analyst, Big Data and Analytics, Wikibon

George Gilbert is big data analyst for Wikibon/theCUBE. He was a big data analyst for Gigaom Research and has been profiled on the front page of *The Wall Street Journal* and published as a guest author in a major overview of cloud computing in *The Economist*. Previously, George was the lead enterprise software analyst for Credit Suisse First Boston, one of the top investment banks to the technology sector, and a product manager on Notes at Lotus Development. George received his B.A. degree in economics from Harvard University.

Twitter | Website | Blog

Download the full e-book: Data Disruption

Historically, analytics and transaction processing have been separate. For example, in call center applications and e-commerce websites, tracking what was happening was very much separate from trying to analyze what should happen or what would be a better outcome. The performance limitations of databases meant you had to change between capturing transactions and analyzing the data. That was slow. When the systems were in operation, data had to be extracted from the transaction systems, transformed into something the analytic system could work with, and then actually loaded into the analytic system. This "pipeline" could introduce a delay anywhere from hours to days or even weeks.

If the analytic system needed new data to answer different questions, the delay was much worse. This took a full development cycle, because someone had to determine in advance which questions to ask. The developers then had to locate those data in the transaction system, transform them into a format that could be used for analysis, and load the resulting data set into analytics systems. The process took too long and was too brittle.

> "The next generation's business applications are going to look a lot more like the leading-edge consumer websites of the past five years—LinkedIn, Netflix, Amazon."

KEY LESSONS

1. WHAT DISTINGUISHES LEADING-EDGE CONSUMER WEBSITES FROM TRADITIONAL APPLICATIONS IS THE INTEGRATION BETWEEN DATA CAPTURE AND ANALYTICS.

2. MAINSTREAM COMPANIES NEED A CUSTOMIZABLE PLATFORM TO TAKE ADVANTAGE OF SUPERFAST DATA COLLECTION AND ANALYSIS BECAUSE THEY DO NOT HAVE THE TECHNOLOGICAL MUSCLE TO BUILD IT IN HOUSE.

THE FUTURE DEMANDS TECHNOLOGICAL MUSCLE

You could not easily change what you wanted to ask, and you could not get the information back when you needed it: the time of the transaction.

In contrast, the next generation's business applications are going to look a lot more like the leading-edge consumer websites of the past five years—LinkedIn, Netflix, Amazon. What distinguishes those sites from traditional business applications is that the analytics and the transaction data capture are integrated. In the case of Netflix, while you are browsing through the catalog one night, Netflix's engine is learning what type of movies you like. The next time you log on, Netflix acknowledges your preferences and either shows a different catalog or organizes the existing catalog differently.

At Wikibon, we call these new applications *systems of intelligence*—systems that learn from past behavior, and then anticipate and influence the consumer in a business-to-business-to-consumer–type application. Think of it like a banking application or a retailer: someone is interacting with the app, but that interaction could be coming through any channel or touch point. At the intersection of all these interactions and at any point in real time we want our app to realize, "This is what the consumer is likely to do or wants to do, and here's how we can best influence them."

> "That type of integration between the transaction and the analytics takes new underlying technology."

One of the first enterprise systems to do that was Harrah's Casino, which implemented a sophisticated loyalty system. Customers were issued a loyalty card, and Harrah's collected data to learn more about customer preferences. If a particular customer had a bad night, Harrah's might offer free tickets to a show it knew the customer liked or provided complementary meals at a restaurant the customer frequented so that the customer would have a better experience and stay longer or stay loyal. For Harrah's, the outcome was better: greater loyalty and profitability related to that customer.

Many organizations want to do the same thing with their business applications, but the Herculean effort that goes into building such websites is not affordable for the "great unwashed"—the Fortune 1,000. The platform technology has to get better so that mere mortals can build these sites.

Most mainstream companies' architectures are based on traditional scale-up Structured Query Language (SQL) databases. Leading-edge consumer websites such as Twitter, Facebook, and Google have much more advanced platform technology. So, we have to make our traditional SQL databases better at this sort of thing. That type of integration between the transaction and the analytics takes new underlying technology. Mainstream customers are going to want to see database providers deliver this as a platform on which applications that can be built and customized.

ABHISHEK MEHTA
CEO,
Tresata

Abhishek Mehta is the founder and CEO of Tresata, a predictive analytics software company redefining business by automating complex human processes. He has built Tresata into a leading analytics innovator with a vision to use data to help enrich life. His history is a combination of radical technology expertise and practical, in-the-trenches business leadership. He was an executive in residence at MIT Media Lab, managing director at Bank of America, and in client-facing leadership positions at Cognizant Technology Solutions and Arthur Andersen.

Twitter | Website

Download the full e-book:
Data Disruption

While attending my 6th consecutive Hadoop Summit earlier this year, looking around to get a read on the pace of innovation in the larger Apache Hadoop/big data space, I was surprised and a little bit disappointed.

Just as in past years, the 2015 conference focused on data infrastructure & transformation—finding new ways to develop clean, organized, formed, and relevant data for decision making. If this is the hot topic even after 6 years, I thought, we are in trouble. Data transformation should be a given, almost a birthright in the new world of Big Data. We need to dream bigger.

KEY LESSONS

1. **THE VALUE FOR LARGE-SCALE TRANSFORMATION ACROSS ALL ENTERPRISES LIES IN AUTOMATING COMPLEX BUSINESS–HUMAN PROCESSES SUCH AS ANTIFRAUD AND ANTI–MONEY LAUNDERING.**
2. ***BATCH* AND *REAL TIME* SOON WILL BECOME OUTDATED TERMS: EVERYTHING WILL BE REAL TIME.**

Here is my advice for moving the collective conversation to the next level:

- Look for areas to automate data management functions that prepare data for monetization. We call this concept the *data factory* and predict that these functions can and inevitably will all be automated.
- Data scientists & engineers must seek to automate complex business–human processes. Examples of these processes - antifraud and anti–money-laundering operations, which are 90 percent manual and 10 percent automated. They should be entirely automatic. The value for large-scale transformation across all enterprises lies in automating these processes.
- Look for ways to combine big data with real-time processing to solve massive business problems. If we do not leverage these capabilities, we will fail.

> “Data transformation should be a given, almost a birthright.”

DREAM BIGGER

My prediction is that we soon will stop using the computing terms *batch* and *real time*. They will be meaningless distinctions. Everything will be real time. After all, the holy trinity of enterprise software solutions—speed, quality, and low cost— unachievable for so long, is finally a reality today. That disrupts everything we know.

The global technology market has been estimated at $3 trillion a year. Roughly 80 percent of that—$2.4 billion—comes from enterprise software. In my opinion, all of it is up for grabs.

Technology is going through a Darwinian era: the enterprise technology space has been built on data stacks which are finally getting decimated. We stack everything—storage, databases, analytical tools, and virtualization tools. For the past 50 years, those data stacks have held all the value.

Stacks have a fundamental problem, however—something I call the "*data hop*" problem. Stacked data must be transported around each tier of the stack. But you cannot move big data. The fastest wires in the world are 10 gigabytes per second. Do the math: moving petabytes (1 Million Gigabytes) becomes herculean. If Moore's law keeps reducing hardware costs and doubling capacity/power, why then can we not improve software to process data as well as store, analyze, and organize data without moving them?

> "The ability to deliver products and services at the right time in the right place to the right customer instantly is the future."

The good news is – we finally can. The answer is Hadoop (and the larger ecosystem associated with it including Spark). We do a big disservice to big data by conceptualizing Hadoop as a storage platform. People do not see it for what it really is: a massively parallel computational engine.

I have called this Big Data era powered by Hadoop the start of the Second Industrial Revolution because data is now the core asset for every enterprise. The ability to deliver products and services at the right time in the right place to the right customer instantly is the future. The technology finally exists to make it happen.

We have to discover and deliver value for each and every customer – whether they are an individual, a small business or a large company. We would then have fundamentally improved how we live. So dream big!

Mighty Guides

Mighty Guides make you stronger.

These authoritative and diverse guides provide a full view of a topic. They help you explore, compare, and contrast a variety of viewpoints so that you can determine what will work best for you. Reading a Mighty Guide is kind of like having your own team of experts. Each heartfelt and sincere piece of advice in this guide sits right next to the contributor's name, biography, and links so that you can learn more about their work. This background information gives you the proper context for each expert's independent perspective.

Credible advice from top experts helps you make strong decisions. Strong decisions make you mighty.