CITO Research
Advancing the craft of technology leadership

OCTOBER 2013

How to Stop Small Thinking from Preventing Big Data Victories

Coauthored by:

Dan Woods
Chief Analyst, CITO Research

Scott Gnau
President, Teradata Labs

Contents

Introduction: Ending Small Thinking about Big Data

It is time to end small thinking about big data

It is time to end small thinking about big data. The perfectly natural excitement about big data and data science has unconsciously narrowed the focus of how to best make use of its potential. Instead of thinking about how to fit the insights of big data and data science into the larger context of business, we often hear simpler questions discussed, such as how to store large amounts of data and how to analyze it in new ways. This thinking is small because it focuses on technology and new forms of data in an isolated and abstract way.

To make the most of big data and data science, we need to start thinking big:

- Big data is really just "data." What's the best technology to handle all of our data?
- Big data provides one piece of a larger puzzle. How can we effectively combine it with existing analytics systems to yield the greatest impact?
- Big data needs to "plug in" and enhance business operations. How can we use big data to create better products and services?

In essence, we need to weave together a new narrative that explains where we started in our effort to support business with technology and data, what progress we have made, and how big data and data science can be added to the mix to create better businesses.

This CITO Research paper weaves a new narrative, defines the principles and patterns of a next-generation data architecture, and explains how and why companies should work on improving how they incorporate big data and data science into existing capabilities.

What's True and False about Big Data

False assumptions about big data contribute to small thinking. On the technology side, because there is a new generation of big data technology, people think a silver bullet will suddenly make data management easy. Or they may think they need to rip and replace existing systems, reinventing the wheel.

Big data is often thought of in isolation, with a focus on special skills. The data scientist is held up as "the sexiest job title of the 21st century," as if businesses need a magical alchemist who can turn lead into gold. Data scientists—and more importantly, data teams—cannot work in isolation, because they must interact with people across the organization to understand the business.

The data scientist is held up as "the sexiest job title of the 21st century," as if businesses need a magical alchemist who can turn lead into gold

The right approach acknowledges that big data and data science efforts are evolutionary and incorporate new technologies and processes into existing ones. It is not about starting over, but about improving current models with new techniques and technologies, enriching those models with an ever-increasing amount of data:

Make no mistake—big data and data science have brought about many developments so far:

- New types of data with varying levels of structure have arrived in high volume.
- Big data has a much lower "signal." The value in sales data is obvious; the value in petabytes of clickstreams is less obvious.
- A new generation of technology applies better algorithms to provide insight into customer behavior patterns.
- New systems, such as Hadoop, have storage and processing capabilities that bring new ways to think about data at scale.
- Faster discovery techniques enable analytical teams to find signals and trends in big data.
- Data scientists and chief data officers combine skills from analytics, software development, statistics, and interaction design.

What Hadoop Is

Given the importance of Hadoop in discussions of big data, it's critical to be clear about what Hadoop is and does.

Hadoop is a file system that is massively parallel and linearly scalable.

Hadoop can store any kind of data without changing the data. In the past, the data stored in Hadoop would have been thrown away because it was too costly to transform it and store it in relational databases. If this data were kept, it would have been summarized in rows and columns. In some cases, Hadoop stores new types of data (mobile GPS coordinates) that didn't exist before.

By storing everything without declaring a schema up front, you don't have to model the data when you store it (that would be inefficient, because much of it may not be used). You can do "on the fly" modeling as you need it, as opposed to modeling everything up front.

By integrating Hadoop with your existing data management and analytic investments, you can use this new granular data to enrich business insights and drive more business value at a lower overall system cost.

No Reset Button: We Are Not Starting From Scratch

These new capabilities don't mean that we must hit the reset button. We still need to harvest information from enterprise applications and construct a comprehensive structured model of our business. We need to securely manage information as an asset. We need to control access to data to protect privacy and comply with regulations. And we need to enable people to explore as much data as possible.

Big data and data science don't mean we flip the off switch on all past business intelligence activities

In other words, big data and data science don't mean we flip the off switch on all past business intelligence (BI) activities. They mean that we understand how to do a better job with everything we have by adding new capabilities.

Few companies are satisfied with how much data they are using. Fewer than 30% of all employees are using BI tools. That leaves a lot of room for improvement and simply starting a big data or data science program will not address the barriers that prevent more people from making better use of data.

To make progress, we should all be trying to answer two questions:

- How can we take the new capabilities and add them to what we already know about creating a next-generation data architecture?
- How can we channel the energy that surrounds big data and data science into achieving a cultural transformation?

Small thinking ignores these questions. We must remember that big data isn't about technology, the three V's (volume, velocity, and variety), a use case, an architecture, or super-genius data scientists crunching numbers. It is much more than that—a movement, a mindset, that's ingrained in an organization.

Using Big Data Energy to Start a Movement

Movements don't succeed without a compelling vision of the future. The goal of a movement catalyzed by big data and data science should be to create a data culture, to build on what we've done in the past, to get everyone involved with data, and to derive more value and analytics from all the data to make business decisions. This is the real victory.

Starting a big data movement involves challenges:

- Transforming company culture to be data-driven and compete on analytics
- Understanding the value in big data and making sure information is integrated across departments/silos
- Ensuring reliable, consistent, and reuse of data across the enterprise
- Discovering nuggets of information about customers, products, and performance across systems and data formats (ERP, legacy systems, web logs, email, voice, text, social media, and more)
- Making data and analytics accessible to as many people as possible

That last point bears repeating. Big data is important, but it is key that businesses use ***all the data*** they have and make it accessible to all its users.

A 10% increase in data accessibility can lead to a $65 million increase in annual net income for the average Fortune 1000 company

Accessibility is crucial: A 10% increase in data accessibility can lead to a $65.67 million increase in annual net income for the average Fortune 1000 company.[1]

In addition, it is important to avoid past mistakes such as failure to engage line of business users, a lack of data governance policies, poor data quality, and failure to validate and test data as it is moved from one system to another.

With so many challenges, creating a clear vision for a movement is clearly a complex design problem. The right vision for each company will differ, but for most companies a movement should be characterized by:

- Using business questions, not technology capabilities, to drive the architecture
- Increasing self-service access to data among users for decision making
- Enhancing intuition and experience with data-driven decisioning
- Enabling fast, iterative discovery that allows analytical teams to "swim" in the data and see what signals or trends emerge
- Improving granularity of and accessibility of customer interaction data to drive sales and improve products and services

The Fundamental Enabler: High-Def Models

The next step in our narrative is to show how big data and data science can actually help improve the way we use data now to drive business performance. This is a key part of the vision that is needed both to support a big data movement and to define a next-generation data architecture.

Without an explanation of the general impact that big data and data science will have on our existing computing infrastructure, we are essentially asking people to figure it out from scratch in each new context.

Fortunately, there is a key capability that shows how big data and data science will make existing applications and analytics systems better: high-definition models.

[1] *http://www.forbes.com/sites/ciocentral/2012/07/09/will-big-data-actually-live-up-to-its-promise/2/*

Remember, most enterprise applications and most business intelligence systems in use today were designed in a time of information scarcity. The whole point of most enterprise applications was to create a system of record that described a business activity with accurate data that was expensive to collect.

Professor Elgar Fleisch, an expert in the Internet of Things, was the first to point out a ground-breaking shift in the economics of data collection. Information, once scarce, is now abundant. Making use of the flood of big data, we can now create high-definition views of our businesses.

Big data does have different properties than traditional data sources. It is often low signal and dirty. However, new analytic techniques can convert copious low-signal data into highly detailed descriptions that enable us to recognize important events.

Just as high-definition broadcasts require different signal processing, so does a high-definition view of big data

Just as high-definition broadcasts require different signal processing, a high-definition view of big data requires new signal processing, including the new styles of analytics that are enabled by MapReduce or by techniques like sessionization, graph, and time series that are algorithmically new (at least to most of the business world). The whole point behind these techniques is to filter out noise, find the signal, and amplify that signal to the point where it's actionable.

Just like a high-definition television, such amplification requires new display technologies as well as new kinds of visualizations to make data meaningful to more people.

The nature of high-definition models is still being understood, but most applications and analytic systems will benefit from a high-definition view. Using high-definition models, customers can be better understood, processes can be analyzed in greater detail, and automatic responses can be programmed. (Consider, for example, real-time auctions for online advertising that place ads based on a customer profile gleaned from many sources including current customer location.)

How High-Def Models Work

A high-definition view requires that *all* the data be exploited to find new and differentiated insights. This means having the ability to save and sift through all the data to deliver an uber-personalized customer experience through micro-segmentation.

Clickstreams are low signal by nature. To derive insight, you need analytical techniques that uncover patterns

For example, thousands of clicks on a clickstream are low signal in nature. Deriving meaning from those clicks requires analytical techniques that uncover patterns. Clickstreams can then be refined into meaningful information and combined with high-signal data that leads to customer or business activity discoveries such as common paths to purchase, product affinity, or shopping cart abandonment.

High-definition models are the foundation for competitive advantage via data-driven apps and predictive capabilities for better customer transparency. Here are some examples of high-definition models in action:

- To increase productivity, a medical supply company combined order data with time/motion studies from sensors on people and warehouse carts, saving thousands of man-hours of labor while relieving congestion in the warehouse.
- A large financial institution synthesized several data sources for a complete, time-ordered view of customer behavior, resulting in a comprehensive customer view across channels, including where customers were coming from and what behavior precedes purchasing new products or churning and closing their account.
- A major telco combined web usage and contact center interactions to identify important new customer churn factors, resulting in identification of hundreds of at-risk churn targets worth millions per year.

An Integrated Way Forward: Next-Gen Data Architecture

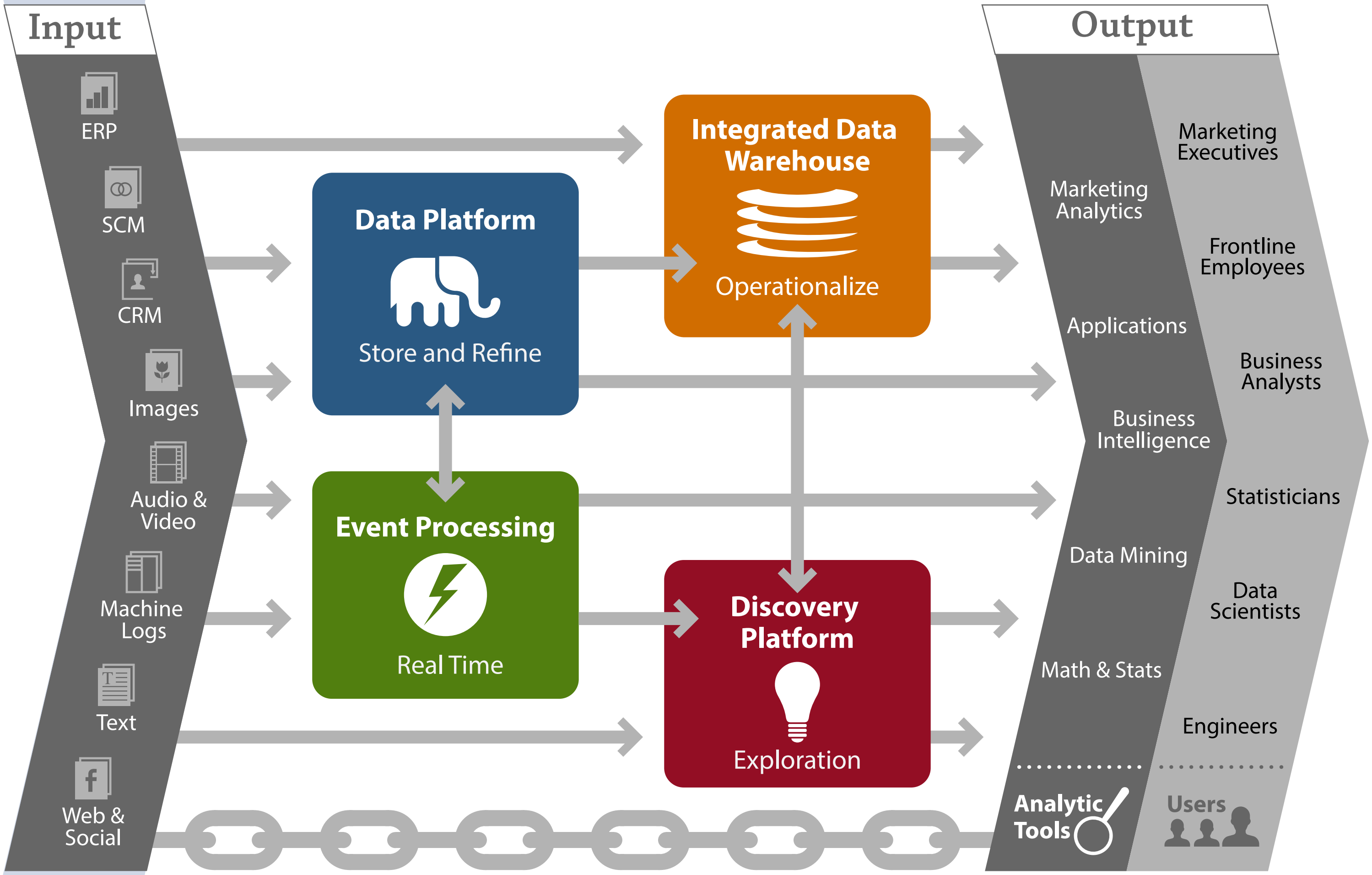
Our next challenge is putting everything together into the big thinking we mentioned at the beginning.

While high-definition models provide an important bridge between the world of big data and existing computing environment, structural enhancements are also needed.

The theory of a next-gen data architecture brings data science, big data, high-definition models, and existing data and technology together, reaching new levels of business value. At the most fundamental level, the next-gen data architecture looks far more like a supply chain than like the model used by most data warehouses.

A next-gen data architecture involves a flow of data to and from many engines. Each engine may create, store, and transform data. Each engine may have one or many jobs. Some nodes support the flow of data, others manage a repository to support a workload, and others organize data from one or many applications, like a data warehouse does.

Next-Gen Architecture: Data Supply Chain



Next-Gen Architecture Capabilities

A next-gen data architecture gives every user and application in the enterprise access to all the data without thinking about where the data is stored or processed

Because of its integration across the company, a next-gen data architecture gives every user and application in the enterprise access to **all the data** without thinking about where the data is stored or processed. It contains the right tools to solve different problems and supports varied data workloads with consistent performance.

Key components like security and data governance are a part of this architecture, just like your existing system, but with even more emphasis on data privacy.

A next-gen data architecture is about getting analytics and BI right and also about adding new dimensions—high-definition granular data, as well as new events—using big data and data science. In addition, it's about analytic innovation. It's not just reporting on the status of the business, but enhancing analytic models and developing new analytic models and algorithms that offer competitive advantage in providing better customer experiences, reducing waste, and stopping fraud.

Teradata has taken the theory of next-gen data architecture and put it into practice with Teradata® Unified Data Architecture™ (UDA).

Principles of a Next-Gen Data Architecture

So if the next-gen data architecture is a data supply chain, what can we say about how it is organized? Of course, the actual structure of next-gen data architecture will be different in each company. But we assert that the most successful implementations will follow these principles:

Principle 1: Open your mind about data. It's all the data, not just big data. There are other kinds of data such as structured, multi-structured, and machine-generated. It's critical to extract value from these data types as well. These new categories of data require new analytic technology while leveraging existing technology investments in a best-of-breed scenario, with less risk and more value. Teradata UDA integrates and analyzes data from every touch point of your organization: transactional systems, web, text, social media, and machine-generated data.

Principle 2: All technology may be relevant. UDA integrates the Teradata data warehouse, the Teradata Aster discovery platform, and Hadoop as key complementary components, along with tight integration with the rest of your enterprise architecture, including event processing, visualization and analytic tools, and more. Of course, the performance demands on these UDA components would demand leading processor technology along with modular storage (such as NetApp E-Series).

Principle 3: Replace the stack with the supply chain. A responsive data system is not designed from the top down like a stack. Rather, it constantly increments and adjusts the flow of data, much like a supply chain. By applying supply chain strategies to your data efforts, you'll know where the data is and how it can be refined and enriched as it flows throughout the supply chain. Because next-gen data architecture transcends the supply chain, providing an overarching view of all the data, integration is fundamental. That's why Teradata refers to this as ***Unified*** Data Architecture.

Don't try to give everyone the same view of data, and don't expect people to bend to the technology

Principle 4: Data users are not all created equal. Different consumers of information have different skill levels, different access rights, and different ways they prefer to receive information. Don't try to give everyone the same view of data, and don't expect people to bend to the technology. You've got to be able to deliver information in the way users want to consume it.

Principle 5: Consider adding a chief data officer. There should be a driving force behind all of the components of this integrated architecture and the valuable data they generate.

Principle 6: Preserve what's working. Leverage the policies already in place around data governance, quality, security, and accessibility.

Principle 7: Use Agile methods. Design small projects with quick turnarounds to get results fast. You can capitalize quickly on what works and not waste time in areas that offer marginal value.

Principle 8: Operationalize your insights. Big data projects can yield exciting results. Unless those results are tied back into operations in a scalable way, the insights you gleaned when working with a segment of your customers can't be rolled out to all your customers. Operations is key to getting real value from all the data.

Next-Gen Data Architecture in Practice

Many types of analyses are thought to be unique to big data, but sentiment analysis, product affinity and market basket analysis, and browsing data analysis were all being used before people knew about big data. Having a high-resolution model of your business doesn't imply a whole new type of analysis. It means that the analyses you've been doing can be fleshed out in new ways, supplemented with new and more information, often at more frequent intervals.

In the past, a company might optimize its supply chain by running truck delivery schedules once a quarter; by adding big data, they can optimize them every day

In the past, a company might optimize its supply chain by running truck delivery schedules once a quarter; by adding big data, they can optimize them every day. In effect, the extra volume of data offers better insight and a more detailed view into operations. Processing the data faster enables you to refresh models when you want.

In place of a solely transactional view of customer activity, behavioral models allow us to sequence customer events and pinpoint predictors. Imagine testing a new product in the summer and inspiring trend setters to tweet about it, then monitoring Twitter for the buzz. That data could then be used to recalibrate production forecasting. This is an example of adding data elements that we couldn't capture before.

Another example is developing churn models. If a caller is angry, the call could be automatically scored based on vocal patterns associated with anger and frustration. Keywords could also be used. If a caller says, "I'm so tired of being ripped off when I travel overseas!" the keyword phrase 'ripped off' could become a new factor for a churn model.

The Ultimate Victory: Starting Your Big Data Movement

Facility with data-driven decisioning should be an important hiring factor for all positions, not just analytic or IT positions

Movements aren't built in a day. So start now, because your competitors probably already have. A worldwide Gartner survey of IT leaders reveals that 64% of respondents had invested in big data technology, or were planning to do so within a year.[2]

With the big data movement, it's about exploiting data, uncovering new insights, and taking action. It's also about creating a big data culture within your organization. Here are some steps you can take.

Have the right people in place. You need data scientists, business analysts, or a combination to crunch the data and find insights. Some companies find that people with IT as well as business experience understand big data technology and business problems. Other companies create a data team, similar to a skunkworks team that operates across business divisions. Facility with data-driven decisioning should be an important hiring factor for all positions, not just analytic or IT positions.

Prepare for stewardship of big data. Oversight of an organization's data assets is critical to giving business users accessible high-quality data. After all, users want data that is useful to them. Stewardship also involves enforcing data usage and security policies.

[2] *http://www.gartner.com/newsroom/id/2593815*

Set big data governance and quality policies. If not governed correctly, big data can run amok in an organization. Data definitions and usage standards need to be set, as does governance for the acquisition, landing, processing requirements, infrastructure management, storage, and security of big data.

Ensure everyone has access to data. There is nothing like promoting a data culture by allowing everyone access to the data. That's the beauty of UDA, which enables transparent access to data for decision making and data-driven applications across the enterprise.

Access and Delivery: The Two Imperatives

The two most important concepts for big data are access to data and delivery of content

Arguably, the two most important concepts for big data are access to data and delivery of content.

Successful access and delivery depend on meaningful integration. From an access perspective, there are knowledge workers, business analysts, and managers who don't know the slightest thing about big data or neural nets, but know how to run a business. Broad access to data is critical.

Once analytics define certain behavior and predict certain outcomes, delivery of that information is also extremely important. When you think about technology for deep data mining in big data, like Hadoop and the extended ecosystem, it does not offer a guaranteed service level for interactive streaming technology. Hadoop is a batch-oriented file system and as such delivers overnight results. Delivery should be consistent, predictable, and sensitive to time constraints.

Since access and delivery are keys to value creation, integration with access and delivery tools is extremely important to the success of your big data movement.

Don't Wait to Start

With big data deployment, actions speak louder than words. Top retailers have seen operating margins increase by 60% through monitoring customers' in-store movements and combining that data with transaction records to determine optimal product placement and mix as well as appropriate pricing.[3] Another fact: The US healthcare industry stands to add $300 billion in revenues by leveraging big data.[4]

The time is now to start a big data movement

That's why CITO Research believes the time is now to start a big data movement. To some it might be a daunting task, but new technologies coupled with existing systems—like Teradata's UDA—solve company problems of access and delivery at scale. With this architecture on the table, there's no reason to wait.

This paper was created by CITO Research and sponsored by Teradata.

CITO Research

CITO Research is a source of news, analysis, research, and knowledge for CIOs, CTOs, and other IT and business professionals. CITO Research engages in a dialogue with its audience to capture technology trends that are harvested, analyzed, and communicated in a sophisticated way to help practitioners solve difficult business problems.

Visit us at http://www.citoresearch.com

About the Contributors

Dan Woods, Chief Analyst, CITO Research

Dan Woods is CTO and founder of CITO Research. He has written or coauthored more than 20 books about business and technology, including APIs: A Strategy Guide. Dan-writes about data science, cloud computing, and IT management in articles, books, and on CITO Research, as well as in his column on Forbes.com.

Scott Gnau, President, Teradata

Scott Gnau is president of Teradata Labs, Teradata's innovation engine. The technologyinnovations from his organization are a driving force behind Teradata's position as the leading analytic data solutions company. Scott is a popular speaker on using data and analytics to create competitive advantage. Working in Teradata's Professional Services business, Scott and his teams drove the successful delivery of enterprise analytics solutions for North America's retailers. Prior to joining Teradata in 1995, Scott owned his own company, focused on data integration and architecture consulting

[3] *http://www.truaxis.com/blog/12764/big-profits-from-big-data/*

[4] *.http://www.information-management.com/news/big-data-ROI-Nucleus-automation-predictive-10022435-1.html*