The Value of Big Data

Using big data to examine and discover the value in data for accurate analytics

Technology White Paper

Third Nature

TABLE OF CONTENTS

Introduction

Data warehousing is a success, judging by its 25 year history of use across all industries. Business intelligence met the needs it was designed for: to give non-technical people within the organization access to important, shared data. The resulting improvements in all aspects of business operations are hard to dispute when compared to the prior era of static batch reporting.

During the same period that data warehousing and BI matured, the automation and instrumenting of almost all processes and activities changed the data landscape in most companies. Where there were only a few applications and minimal monitoring 25 years ago, there is ubiquitous computing and data available about every activity today.

Data warehouses have not been able to keep up with business demands for new sources of information, new types of data, more complex analysis and greater speed. Companies can put this data to use in countless ways, but for most it remains uncollected or unused, locked away in silos within IT.

There has been a gradual maturing of data use in organizations. In the early days of BI it was enough to provide access to core financial and customer transactions. Better access enabled process changes, and these led to the need for more data and more varied uses of information.

These changes put increasing strain on information processing and delivery capabilities that were designed under assumptions of stability and common use. Most companies now have a backlog of new data and analysis requests that BI groups are struggling to meet.

Enter big data. Big data is not simply about growing data volumes — it's also about the fact that the data being collected today is different in ways that make it unwieldy for conventional databases and BI tools.

Big data is also about new technologies that were developed to support the storage, retrieval and processing of this new data. The technologies originated in the world of web applications and internet-based companies, but they are now spreading into enterprise applications of all sorts.

New technology coupled with new data enables new practices like real-time monitoring of operations across retail channels, supply chain practices at finer grain and faster speed, and analysis of customers at the level of individual activities and behaviors.

Until recently, large scale data collection and analysis capabilities like these would have required a Wal-Mart sized investment, limiting them to large organizations. These capabilities are now available to all, regardless of company size or budget. This is creating a rush to adopt big data technologies.

As the use of big data grows, the need for data management will grow. Many organizations already struggle to manage existing data. Big data adds complexity, which will only increase the challenge. The combination of new data and new technology requires new data management capabilities and processes to capture the promised long-term value.

Understanding Big Data

The value of technology

To understand the value of novel new technologies you need to differentiate between two things that are often confused by analysts and technology media: capabilities and functions.

Capabilities derive from combinations of functions. Functions are the basic tasks or activities that can be performed with the new technology. Broadly speaking, a capability is what you can achieve with the technology: what it's *for*. A function is what the technology *does*.

For example, a lamp does one thing: it gives off light. There are different lamp technologies: electric lamps, gas lamps, oil lamps. They all perform the same function, but with different constraints and features.

The lamp's primary function is to give off light, but it enabled a new set of activities that were previously costly or difficult. The new capability delivered was the ability to remain active after dark. Prior to inexpensive, ubiquitous lighting, even basic tasks like cooking and eating meals were normally done during daylight hours. Reading by oil or candle light was difficult and more costly than most people could afford. New technologies like gas lamps reduced the cost of lighting and enabled more activities, although most industrial production work was still done in daylight. The advent of electric lighting changed all of this.

Information processing technologies offer an interesting parallel. They function primarily as an enabler of other activities. Each new generation of technologies removes prior constraints or makes a different set tradeoffs to support some new uses at the expense of others.

Big data capabilities

If we look at Hadoop, one of the primary technologies in the big data market, we see that it has three functions: it stores data, it retrieves data and it processes data.

Does it do this like a relational database? No, it stores files, objects or records. It retrieves data, but has no built-in facility for relating one piece of data to another, which is an inherent element in relational databases. Likewise, it processes data, but not in a query that joins and summarizes. It processes data via programs written by developers or generated by tools built for that purpose.

Big data encompasses a set of technologies (like Hadoop, NoSQL databases and real-time processing frameworks) that embody assumptions and design constraints different from what is standard in the database and business intelligence (BI) market. The design and architecture of these technologies offers a combination of four technical capabilities:

1. *Scale on two axes: storage and processing.* Big data technologies have the ability to store and access almost any volume of data. They also provide processing facilities over this volume of data. Scalability can be dynamic, growing and shrinking as processing needs change. Analytic databases are able to scale on the storage and query axis, but have fallen short when it comes to processing large data volumes.

2. *New data accessibility.* Unlike a relational database, where any new information must be conformed (usually via ETL tools) into a model that is built in advance, big data technologies relax the data type and model constraints. A database may not be capable of storing some types of information, but anything digital can be treated as data by a big data platform, from traditional transaction records to digital media.

3. *Highly scalable real-time processing.* The ability to store data as fast as it can be produced is a key feature of big data storage technologies. Some of them offer very low latency processing or retrieval, which is how highly scalable interactive web applications are built. Real-time isn't limited to storage and retrieval - it's also possible to process, query and monitor real-time data as it streams across networks. Real-time is an area where databases have not kept pace with business needs.

4. *Ability to run arbitrarily complex analytics.* The processing technologies are not constrained by database interfaces and scale or cost-of-scale limitations. More types of analytics are possible, on more data and with greater frequency. We've had batch and ad-hoc analytic tools for years, but generally not in a single platform that combined scalable storage and processing, and not in one that permitted arbitrary code to run over the data.

These technical capabilities are key to understanding what can be done with big data. The difference between big data and traditional models for data processing is the architecture — big data solves them all with one platform, where conventional approaches require multiple systems with different and sometimes incompatible architectures.

The new capabilities are achieved at a lower price than in the data warehouse environment. This doesn't mean big data technologies are inexpensive, only that they are less expensive than what is available in the traditional BI market to achieve the same goals.

Relational databases and multi-processor Unix servers had similar capability and cost impacts on reporting systems and pre-relational databases in the mainframe and mini-computer market of the late 1980s and early 1990s. Business intelligence was the big data of that period.

Like the early gas lamps, BI enabled new activities and practices, bringing illumination and improved transparency to business decision-making. Prior to the advent of BI, decision-making was often done in the dark; BI helped bring it into the light. But even post-BI, business decision-making is still a twilight activity. The constraints of the data warehouse model are like the limitations of gas lighting. One could only shed light near gas mains, and each new light required carefully piping the gas to a fixture.

Electric lights changed this. Bright, non-flickering lights could be used. More importantly, it was far simpler to string electric wire to where the light was needed. It's in this respect that big data enables a new kind of illumination. As with the advent of electric lighting, big data has the potential to eradicate limitations that constrain the effectiveness of the data warehouse-driven BI model.

Is new technology needed?

It's reasonable to ask if a new set of technologies is required to gain these new technical capabilities. The truth is that there are overlaps between what the current data warehousing and BI environments can do and what big data environments can do.

Data warehouses were designed with a particular set of assumptions: that we could gather the data requirements in advance, that we could be aware of and examine all of the data prior to loading it, and that we would consume data as-is from the warehouse.

These design assumptions work well for the shared data used to monitor daily business operations. They do not work well when the need is less predictable, poorly defined, or one involving low latency of data. A core belief underlying the data warehouse is that data requirements do not change significantly or frequently.

Inability to adapt easily to change is the origin of many complaints about BI. The systems aren't flexible and can't react quickly enough to changing information requirements. The result for IT is a backlog of work to capture, integrate and store new data, update BI models, and deliver information to the users.

BI methodologies build what is essentially a publishing model for data: all the work is done in advance to prepare data for read-only consumption. The structure for storing the data is fixed, like chapters in a book. A user can only access the data. There is no ability to refine or combine data and save the results, short of exporting it to Excel or some other tool and working on it there.

Not only is the BI publishing model archaic, the context in which the BI model expects information to be consumed is similarly antiquated. It's like reading a book by lamplight or candlelight – what used to be called "elucubration." While universally practiced, elucubration was never ideal because of the constraints of dim, flickering flames as lighting. With the advent of incandescent lighting it became as easy to read in the dark of night as during the light of day.

There's a sense in which consuming BI is like elucubrating: instead of straining to read or to interpret by the flickering half-light of a candle, business decision-makers are making decisions in the half-light of a computer-screen. They're guessing or interpolating information to plug the gaps in BI's static publishing model.

Big data, like electric lighting, illuminates previously unlit corners. It delivers both brighter lights and the ability to have them when needed. Instead of waiting months for data to be perfectly clean and ready for use, it's possible to use big data technologies to examine and discover the value in data. When valuable, the data can be sent through the more rigorous processes for a data warehouse.

Big data augments business intelligence

Big data technologies arose from different needs than those that drove data warehouses, and are based on different assumptions. The primary need was cost-effective storage and processing at very large scale. This could only be accomplished by using a hardware layer built from many independent servers, allowing for easy growth by adding more servers.

As with the early days of other enterprise software, the first big data systems were built for operational applications. The initial use cases were fast interactive writing or reading by web applications, batch data processing and batch analytic processing.

The new application environments are different from the previous client-server generation. They were developed assuming cheap and scalable hardware, using agile development models that required rapid evolution of code, more developer control, and therefore flexibility in the underlying data storage and retrieval systems. Because of the data processing roots, big data technologies are suitable for uses beyond what they were originally designed for.

Unlike the data warehouse, they don't assume that it is possible to define and model all data requirements in advance. Instead, they expect changes to data sources and targets, and provide features for flexible storage and access of data, as well as more developer control over the environment.

Another changed assumption is the read-only model of the data warehouse. Big data environments were built by and for developers, assuming that data would be read and written. The result is that new tools for data exploration and analysis can be built to overcome the limits of read-only reporting and dashboard tools.

The big data stack is a data processing platform, not a query platform. It combines elements of databases, data integration tools and parallel coding environments into a new and interesting mix. The IT market today distorts this view by looking at big data as a replacement for one of these technologies.

Looking at big data as a replacement for existing applications overestimates the potential for displacement of products while underestimating the impact it has on the IT architecture these products operate in. Big data augments business intelligence, it doesn't replace the BI stack.

The value of big data

Big data is a term used as a catch-all for data, technology and methods for the processing of information, just as business intelligence is a catch-all term for data, technology and methods for query and retrieval of information.

The business value of big data is in how the data and technology are applied, the same as with BI. Big data's value derives from capabilities that enable new and broader uses of information, or that remove limitations in the current environment.

Unlimited scale of storage and processing provides new flexibility

Query scalability has been solved by multi-node parallel databases, some of which are in the petabyte scale today. Many data warehouses have been built using these databases, leading people to question the need for big data technologies.

One benefit big data delivers is agility, lost in the highly controlled data warehouse environment. Adding a new data source to a warehouse is a slow process. Each new dataset involves multiple roles: an analyst to gather requirements, a data architect to model the data, a database administrator to make database changes, an integration developer to process and load the data and a BI developer to make the data available.

It is not always necessary to gather requirements, model the data in advance and do robust data cleansing and integration. Instead, store the new data in its raw form and give access to it as-is. Users and developers can access and transform data that doesn't fit in a highly controlled data management environment. This creates a less restrictive data layer separate from the data warehouse, enabling uses that are not well served today.

Big data systems do not generally require or enforce strict controls over the model in which data is stored. Instead of processing data in advance, the data can be processed at access time. This gives rise to more flexible methods to manage and process data.

Think back to the constraints imposed by different lamp technologies. Prior to the invention of electric lighting, it was impossible to use lamps, candles, or even gas-lights in a wide variety of use-cases – including many industrial applications. So it is with BI, which tends to emphasize rigidity and strict control because of the constraints imposed by its architecture. As is the case with lamp technologies, there are applications where using the data warehouse would be counter-productive, if not destructive.

Big data processing isn't without constraints of its own. For the most part, however, its capabilities address gaps or shortcomings in the BI stack. The big data approach works particularly well for one-off requests and data of unknown long term value — problem areas for BI. Simple user requests for new data that would have gone unserved can be met with the new platforms, albeit with fewer guarantees about quality.

Loosening restrictions on what data is loaded allows one to store and use a broader set of information. Data warehouses exclude data of unknown value due to inability to justify the work, or the effect it has on size and therefore query performance. Many organizations archive data from BI environments for the same reason, or because retaining history increases size which in turn requires more hardware and software licenses, driving costs higher.

Archiving data due to cost and performance reasons makes a tradeoff between business value of data versus the IT costs of keeping it. Big data changes the cost and utility equation, allowing for online retention of *all* the data desired, for as long as desired.

New data accessibility

Big data usually enters when the problem involves more than reporting and query performance. If the need is for processing rather than simply querying a large volume of data, or the need is use of data that doesn't fit in a relational model, the database-oriented warehouse model breaks down.

Combining scalable processing with flexible data structures means that any digital content can be treated as data. This opens up entirely new sets of data to query, analysis and use by the organization. Internal content from customer facing systems, streams of events, content repositories, even text from external sources, can be processed and used.

There are two usage models for big data in this context. One is using a big data platform to search for, browse and analyze complex data or content to gain insight; the discovery or "data science" uses.

Databases have trouble managing data that isn't in orderly rows and columns. Complex data with hierarchies, nested elements, implicit ordering or networks of relationships doesn't fit well into the schema model and SQL access of a standard database.

The poor fit means specialized, purpose-built software is required. Traditionally, this would have meant a separate application to store, process and analyze each type of data, for example one for text and another for network data. Big data technologies can store and process any kind of data, enabling support for all of it.

SQL is the only means of access in a database. This is not a restriction with big data, which means analysts have new modes of interaction. They can use search to find information, other tools to browse data and content to discover new links, and analyze the data with other techniques, all on the same platform.

The second usage model is big data as a data integration platform for complex data. Content and complex data can be processed to extract data or derive information. This information can then be structured and loaded into downstream databases for use via traditional BI tools.

By combining these two uses, it's possible for an analyst to gain new insights and turn them into a new source of data that is loaded into the data warehouse. In this way, big data bridges the world of complex data and the structured world of the data warehouse.

Scalable real-time processing

Big data isn't just about collecting and processing large amounts of data. It also encompasses new technologies to create highly scalable, interactive, customer-facing applications, monitor and act on real-time data without a human in the loop, and add analytics into existing applications.

Real-time support is rare in the data warehouse world because it is difficult, even at moderate scale. The data must be stored before it can be accessed, and the only access mechanism is to query the database. The result is that monitoring data and taking action is slower than operating on it as it moves. Performance tuning the simultaneous loading and BI workload is hard, to the point that many can't do it.

There are multiple big data solutions to handle high volume streaming data in real time. These vary based on whether the problem being solved is real-time ingest and processing of data for later use, real-time retrieval of data, or real-time in-stream processing to monitor, deploy analytic models or trigger actions.

Real-time analytics combines the scalability of batch processing on a big data platform with the ability to act on streaming data. Analytic models are built using the large amounts of data collected. The models are then deployed in a different layer that executes them on the interactive or streaming data.

The combination of monitoring, processing and storing data in real-time and at large scale is a use case that falls well outside the design or capabilities of most data warehouse platforms. This is a high value area for the deployment of big data solutions for two reasons: it's been extremely difficult, and there is a shift in almost all industries to lower latency communication and coordination between companies and customers.

Analytics

Advanced methods of analysis have been available in the BI market for a long time. For many companies, analytics were previously out of reach or in limited use due to the software and processing costs, particularly at large scale. Building and running a single analytic model is both data- and process-intensive enough to bring a data warehouse to a standstill, stopping all BI work.

Analytic workloads are different from BI. They require extensive data transformation prior to model execution, regardless of the source of data. This is one reason analytics is presented as a killer app for big data. It's possible to store, process and analyze data in a big data environment more easily than in the traditional database-oriented warehouse. The big data platforms were originally designed to support systems that process data to produce results, not query data to deliver an answer.

Big data removes the technical limitations while adding to the creation and availability of ever larger volumes of data. By providing scalable processing that can grow and shrink dynamically, the new platform changes business analytics in three ways:

> ***More analysis.*** When scalability is not a problem, models can be run in minutes instead of days, which means they can be run more often. For example, supply chain analytics are often run weekly due to the cost and difficulty of scaling them. Several retailers have shifted their optimization models to run daily or at the end of each shift, enabling major changes to operations. Analytics can also be applied to more areas of the business, where before it had been reserved for a handful of high-value problems.
>
> ***Deeper analysis.*** The availability of new data and scalable processing makes it possible to construct models that take into account more information. Cheap scalability opens the door to new algorithms that weren't used due to their unacceptable performance, and new algorithms that need more or different data than was previously available.
>
> ***More broadly accessible analytics.*** A big data platform makes it possible to easily combine analytic models with the nightly processing of data into the data warehouse. The platform makes it equally easy to process data and make it available to operational applications. Adding real-time technology means analytics can be integrated with interactive applications or run live on streaming data.

A key element big data provides for analytics is the use of a single platform for processing. In a traditional warehouse architecture, data is moved to another server, analytic models are run, and the results are then delivered to users from there. A big data architecture simplifies the environment. It stores, transforms and processes the data without multiple tools and without moving data between siloed systems.

Using big data technologies, one developer can do all the work within a single environment without the need to involve system and database administrators, speeding up development. A popular use of big data platforms is to serve as the processing engine for text and other data incompatible with the tabular formats of a database. Once built, the processing and analytic models can send the output to a data warehouse, where it is more easily delivered to users.

Big data is doing for data processing today what the data warehouse did for query and reporting in the early 1990s. The innovation of the BI publishing model brought new

illumination and transparency to decision-making and business operations. Even with the unprecedented illumination of BI's "lamp," there are still many dark areas.

The activities and practices enabled by big data – some of which are long-sought, others of which are previously unimagined – have the potential to illuminate regions or spaces of business decision-making and business operations that will *always* be dark to BI. The technology shift has broad applicability, particularly in industries with large data volumes or complex data analysis requirements, such as the retail sector.

A look inside the evolving information needs of the retail sector

Effective consumer product development and retailing is about differentiation. The primary areas of differentiation today being customer experience, branding and service. Investment priorities should be in customer facing areas since these drive retention and growth. This means building new analytic capabilities using a repository of business information beyond core sales transactions that are the usual focus of analysis.

Most measurements in retail today are designed around the "average customer" and individual channels. The BI metrics in use are simple ones like average basket size, average items per order or profit per customer. Measurements like these ignore the individual behaviors that shape customer response to products, promotions and services.

Planning around a non-existent "average customer" ignores what we know about how customers drive profitability. Effective retail requires more sophisticated customer analysis strategies to understand behavior across a growing set of channels.

The new analytic requirements are turning top-down analysis on product sales into a bottom-up exercise of understanding individual behaviors. Planning product features and services should be driven by these details, requiring changes to planning and analysis. Both manufacturer and retailer marketing and merchandising optimization need to incorporate the behavioral differences of customers across channels.

To enable detailed analysis requires the integration of data from each point of customer interaction, whether it is marketing, a transaction or a service. This interaction data must be married with all of the existing transaction data as well as data from loyalty programs, social media and other external sources.

New analysis capabilities require changes to the existing BI infrastructure. More data, increasingly complex analysis and the need for up-to-date information stress the current systems. Big data platforms offer opportunities to store and process this information, expanding visibility across channels and into customer behaviors.

For more reading related to understanding big data technologies, we have also written the following white paper, "*The Challenges of Big Data & Approaches to Data Management*". In this paper we discuss two areas driving big data use in retail and CPG today: multi-channel retail and understanding customer behavior, with a focus on big data challenges and approaches to data management.

About the Author

Mark Madsen is a research analyst focused on analytics, big data and information management. Mark is an award-winning architect and former CTO whose work has been featured in numerous industry publications. He is an international speaker and author. For more information, or to contact Mark, visit http://ThirdNature.net.

Third Nature is a research and consulting firm focused on new practices and emerging technology for business intelligence, analytics and information management. The goal of the company is to help organizations learn how to take advantage of new information-driven management practices and applications. We offer consulting, education and research services to support business and IT organizations and technology vendors.

About SAP

Master Big Data with SAP Information Management Solutions

Information management (IM) is the practice of managing data used by operational applications and analytic solutions to support day-to-day operation and decision-making processes. It elevates raw data into valuable information that can help drive operational excellence and competitive advantage.

Your organization can maximize its return on big data analytics by using Information Management solutions from SAP to gain a complete view of information by accessing and integrating data from any scale from any data source with high velocity. Get unprecedented insight from Big Data by extracting useful intelligence from unstructured data and combine it with structured data for new contextual insight. Ensure trust in information by governing data quality, correcting issues during data movement, and defining policies to know when data is fit for use.

For more information, visit www.sap.com/eim.

Third Nature, Inc.
PO Box 1166
Rogue River, OR 97537