The Big Data Zoo—Taming the Beasts

The need for an integrated platform for enterprise information

October 2012

A White Paper by

Dr. Barry Devlin, 9sight Consulting
barry@9sight.com

Big data is probably the single most important trend in information usage for both business and IT in the past decade. It is changing the way companies make decisions, do business, succeed or fail. Using information and insights intelligently to anticipate and profit from change. It is causing IT to look beyond traditional technologies, to new tools to process larger data volumes of a variety of types faster than ever before required.

Much of the focus on the big data zoo has missed one key point: big or small, it's still data. It must be managed and integrated across the entire enterprise to extract its full value, to ensure its consistent use. Taming big data is the key to gaining that value. This paper offers three key take aways:

1. ***The foundation for extracting the maximum business value from big data at its source is a technically diverse and deeply integrated platform for all information, both big data and traditional transactions***
2. ***An enterprise-level approach—platform, products and processes—is mandatory to ensure long-term quality and use of big data in concert with existing data from business intelligence and other systems***
3. ***Rapid deployment of big data projects is needed to take earliest advantage of emerging business opportunities and is achieved by introducing big data capabilities incrementally into the current data management framework based around data warehousing***

Taming the big data beasts is the next big step in data management.

Contents

Sponsored by:
International Business Machines
www.ibm.com

If big data were a mammal, it might be an elephant. You may be thinking of a small, yellow elephant. But, I'm not; I'm thinking of the large, grey kind—in a herd. I mean more than Hadoop. I mean all of the different types of data that businesses collect in ever-growing quantities. Big data in this sense, despite its technical novelty, is like all other data: it must be properly managed and used at the enterprise level to deliver the significant business value and long-lasting impact it promises.

If big data were a reptile, it might be a chameleon. Before 2005, big data was a phrase used in fear by scientists who couldn't afford to store or analyze all the data their experiments produced. It still is. Next, big data became mostly a playground for researchers at companies like Google and Netflix using the enormous amounts of Web-sourced data they held. It still is. In 2008, Hadoop became a top-level, Apache open-source project and became synonymous with big data. It still is. By 2010, even the Economist had a special report[1] on big data; and marketers began relabeling everything in sight. It still is all of these things… and more. But now, it's time to stop the shape-shifting. Now, big data is the focus of in-depth, advanced, game-changing business analytics, at such scale and speed that the old approach of copying and cleansing all of it into a data warehouse is no longer appropriate. Much of that analysis has to be performed on big data in its native format as close to its source as possible. And a little thought leads directly to the conclusion that a federated or virtualized approach—bridging the data warehousing and big data—will be required.

For business, big data offers new and vital analytic and predictive opportunities enabling them to significantly outperform their competitors[2]. In 2011, McKinsey estimated that big data could generate $300 billion of value in US healthcare and $250 billion in EU public sector administration[3]. Significant business opportunities clearly exist and early movers are already taking advantage. However, all is not rosy in the garden. Among Gartner's predictions for 2012 and beyond[4] was: *"Through 2015, more than 85 percent of Fortune 500 organizations will fail to effectively exploit big data for competitive advantage"* because of their inability to deal with the technical and management challenges associated with big data.

Big data offers big opportunities, but few companies will exploit them effectively.

These challenges are aimed squarely at IT. Big data does not stand alone in the infrastructure. To deploy and use it effectively, it must be embedded in existing business processes. It must and will sit with all the information-centric tools currently in use in a comprehensive, enterprise-scale platform. Big data is best introduced incrementally and, in many cases, sooner rather than later for maximum benefit.

But still, the three big illusions of big data remain. First, it sounds like it can solve world hunger—or, at least, make every business instantly successful. Second, it appears to displace all traditional business intelligence and data warehousing. Third, it seems like everybody is doing it. Unfortunately, for all three… NOT!

So, it may seem that big data is more like a virus, evolving and exploding like some pandemic? The truth is both more revolutionary and more mundane. Big data can and will unleash radical business opportunities… but, only if we return to our roots of good data management processes and well-integrated, enterprise-grade technologies.

Of blind men and an elephant—seeing the big data picture

"It was six men of Indostan / To learning much inclined,
Who went to see the Elephant / (Though all of them were blind),
That each by observation / Might satisfy his mind" [5]

The story of the blind men who feel different parts of an elephant and come to blows over what it looks like—or learn that all truth is relative—describes well the current market situation around big data. Every consultant and vendor sees and describes big data through the parts they touch… not to mention the tools they have and the markets to which they aspire. The overall outcome is confusion… which we'll clear up right away.

Big data is growing rapidly, but defining it precisely is a challenge.

The amount of information stored and processed is growing at over 50% compound annual growth rate according to IDC[6]. This characteristic is called, reasonably enough, ***volume*** and leads to the big data moniker. Most definitions of big data add another two v-words: ***velocity***—the increasing speed of data arrival and processing—and ***variety***—the widening range of data structures that need to be handled. IBM has recently introduced a fourth—***veracity***—the need to trust the data used to make strategic and operational decisions. Some analysts add further v-words: variability is often added and value, virality, validity and viscosity are among other contenders. The definitions are neither convincing nor consistent. In truth, *vague* is probably the most appropriate v-word to use—none of them is amenable to a specific measure. So, how can ordinary mortals answer the simple question: does big data apply to me?

The simplest approach, driven pragmatically by what early adopters in the market have done, is to look at the business uses of big data and see how they apply to you. Of course, this approach cannot be complete, because more novel uses will likely be found. However, big data is likely to be important if your business is following one or more of the following directions:

*Volume, velocity and variety matter far less than what you **do** with big data.*

1. ***Marketing*** uses social media content and relationship information as well as internally captured content from customer interactions such as call center logs to more deeply understand customer motivation. In industries such as retail, consumer packaged goods and telecommunications where there is direct or indirect interaction with large numbers of consumers, this enables a move from sampling to full dataset analysis, from demographic segments to markets-of-one and from longer-term trending of historical data to near real-time reaction to emerging events. The ultimate goal is prediction of customer behaviors and outcomes of proposed actions such as next best offer.
2. ***Detecting fraud*** and other irregularities in financial transaction data has expanded to include larger volumes of often smaller-value transactions on ever-shorter timescales. Big data analysis techniques on streaming data—before or without storing it on disk—have become the norm.
3. ***Real-time forecasting*** becomes possible as utilities, such as water and electricity supply and telecommunications, move from measuring consumption on a macro- to a micro-scale using pervasive sensor technology and big data processes to handle it. Value arises as consumption peaks and troughs can be predicted and, in some cases, smoothed by influencing consumer behavior.
4. ***Tracking of physical items*** by manufacturers, producers and distributors—everything from food items to household appliances and from parcel post to container shipping—through distribution, use and even disposal allows business process optimization or improved customer experiences. People, as physical entities, are also subject to tracking for business reasons or for surveillance.

5. ***Reinventing business processes*** through innovative use of sensor-generated data offers the possibility of reconstructing entire industries. Automobile insurance, for example, can set premiums based on actual behavior rather than statistically averaged risk. The availability of individual genomic data and electronic medical records presents the medical and health insurance industries with significant opportunities, not to mention ethical dilemmas.

There are also some common misconceptions—perhaps arising because some so-called experts are focusing too closely on individual parts of the elephant—that must be addressed.

Big data is much more than just social media feeds from the likes of Twitter and Facebook. This type of data is important, but mostly in the context of the real customers and actual business transactions that we traditionally record in operational systems and measure in business intelligence (BI). Similarly, a single-minded focus on sensor data coming from the growing "Internet of Things" misses the point that use or analysis of such data must somehow fit into existing or reinvented business processes. Nor can an army of disconnected data scientists, however large, hope to effect business change by playing with a single source of data on a new technical platform. Integration of data from a variety of sources, both traditional and new, with multiple tools, is the first prerequisite. A well-integrated process around all data, both big and small, is a further necessity to extract business value from information.

Then, there exists the thought that big data technology can or should displace relational databases or enterprise data warehouses (EDW). This is simplistic in the extreme. In fact, big data technology is actually an extension and integration of existing tools and techniques from batch processing to database management systems. The Hadoop ecosystem, for example, is basically a system for parallel processing of large files in batch. Relational databases and supporting tooling focus on systematic information management, data consistency and more. Big data technology, conversely, emphasizes other desirable characteristics such as speed of access, schema variability and, of course, enormous data sizes.

The truth today is that many leading-edge business processes need both sets of characteristics. Some tasks require flexibility, loose boundaries and innovative approaches; others need certainty, limited scope and adherence to rules. Business processes are crossing a threshold of complexity that is beyond the capabilities of the highly regulated data processing of traditional systems, but is equally unsupported by the simplistic view of a big data world characterized by volume, variety and velocity. We need an enterprise-grade platform and tool set that supports both.

Future-proof business processes require both big and traditional small data approaches and tools.

To define such a platform, we must recognize that we are moving rapidly from a world where one type of data reigned supreme to one where three distinct types of information contribute equally.

Big data and small—a bird's eye view

"The whole wide ether is the eagle's way:
The whole earth is a brave man's fatherland"[7]

Envisioning IT as an eagle, soaring above all of the divisions and silos of existing systems and organizations, we begin to see how all the information and processes interrelate. At a fundamental level, we need a new mental picture of the information landscape, and its three distinct, but deeply interrelated, domains:

1. ***Human-sourced information**[*]:* People are the ultimate source of all information. This is our highly subjective record of our personal experiences. Previously recorded in books and works of art, and later in photographs, audio and video recordings, human-sourced information is now largely digitized and electronically stored everywhere from tweets to movies. This information is loosely structured, ungoverned and may not even be a reliable representation of "reality", especially for business. Structuring and standardization—for example, modeling—is required to define a common version of the truth. We convert human-sourced information to process-mediated data in a variety of ways, the most basic of which is data entry in systems of record.

2. ***Process-mediated data:*** Every business and organization is run according to processes, which, among other things, record and monitor business events of interest, such as registering a customer, manufacturing a product, or taking an order. This data includes transactions, reference tables and relationships, as well as the metadata that sets its context, all in a highly structured form. Traditionally, process-mediated data formed the vast majority of what IT managed and processed, including both operational and BI data. Its highly structured and regulated form makes it ideal for performing information management, maintaining data quality and so on.

3. ***Machine-generated data:*** We have become increasingly dependent on machines to measure and record the events and situations we experience physically. Machine-generated data is the well-structured output of machines—from simple sensor records to complex computer logs—considered to be a highly reliable representation of reality. It is an increasingly important component of the information stored and processed by many businesses. Its volumes are growing as sensors proliferate and, although its structured nature is well-suited to computer processing, its size and speed is often beyond traditional approaches—such as the EDW—to handling process-mediated data.

The ultimate source of traditional business data is personal experiences and machine measures; big data thus directly reconnects business processes to the reality of the world.

The relative sizes and perceived importance of these three domains has shifted over the past decade and is likely to shift further in the coming one. Up until the end of the last millennium, process-mediated data was dominant; the human-sourced information and machine-generated data that existed in digital form was relatively small in volume and considered unimportant in comparison to the well-managed data in operational and informational systems. The last decade or so has seen an explosion of big data, consisting of both human-sourced information and machine-generated data; the former, in the form of social media data, has captured the limelight. The rapid growth of the Internet of Things will propel machine-generated data to highest volume and importance in the coming years.

[*] In the context of these three domains, I use "data" to signify well-structured and/or modeled and "information" as more loosely structured and human-centric.

However, as seen in figure 1, human-sourced information and machine-generated data are the ultimate sources of the process-mediated data on which we have long focused, although only a small and well-defined subset moves through the traditional business process layer that intervenes. These sources are both more flexible and timelier than traditional process-mediated data. In fact, the business processes that create process-mediated data are designed to reduce flexibility and timeliness in order to ensure the quality and consistency of the resulting data. This is most clearly seen in the processes that populate the EDW, but also occurs in operational systems where data cleansing and validation processes ensure the veracity and viability of the data thus recorded.

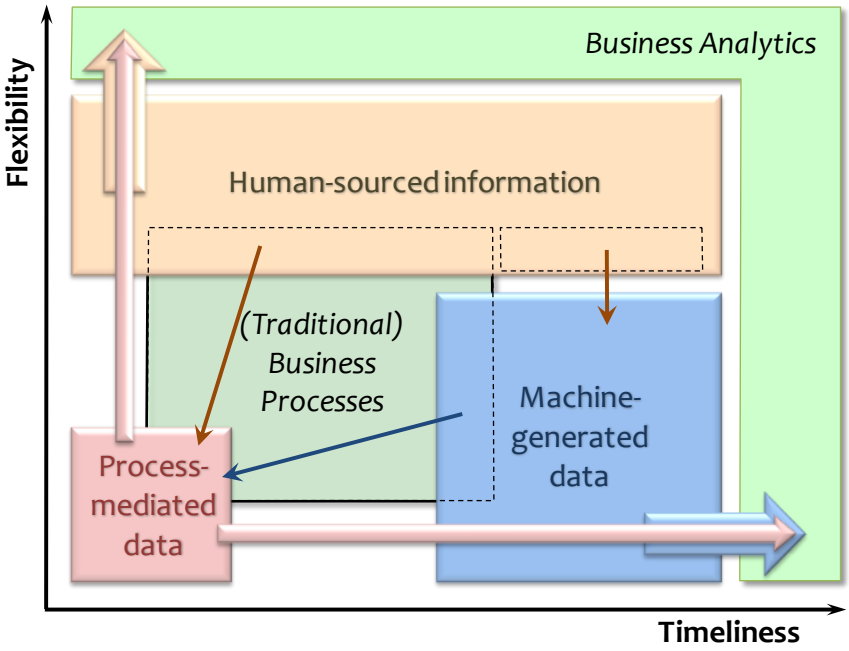


Figure 1: *The three domains of information*

The volumes of human-sourced information and machine-generated data are now much larger and their rates of change and variability higher than process-mediated data. Copying and transforming them to the traditional process-mediated domain is increasingly impractical. Therefore, specialized technology—***business analytics***—is often required to process and explore both human-sourced information and machine-generated data as close to their sources and as quickly as possible. But, of equal importance is the flow of process-mediated data and associated metadata into the business analytics environment to create meaning, context and coherence in the analytics process. Big data and business analytics, in essence, complete the closed-loop information process that has always been implicit in IT.

Business analytics processes big data as close to its source as possible for maximum speed and efficiency.

The practical implications of this three-fold information model are significant and wide-ranging:

- Big data processing, whatever the technology used, depends on traditional, process-mediated data and metadata to create the context and consistency needed for full, meaningful use
- The results of big data processing must be fed back into traditional business processes to enable change and evolution of the business
- A fully coherent environment, including an integrated platform and enterprise-scale organization are necessary for a successful implementation

As big data becomes ever more prevalent, the challenge for business and IT is to move from their previous complete dependence on process-mediated data and embrace these more fluid and changeable classes of information about the real world. Understanding and working with the relationship between the three information domains is fundamental to using big data safely and productively within the business. Defining and managing this relationship and making all three types of information equally and safely available to the business requires an ***integrated information platform***, which is the topic of the next section.

Traditional process-mediated data and metadata is vital to understanding the context and managing the use of big data.

An integrated platform for all information types

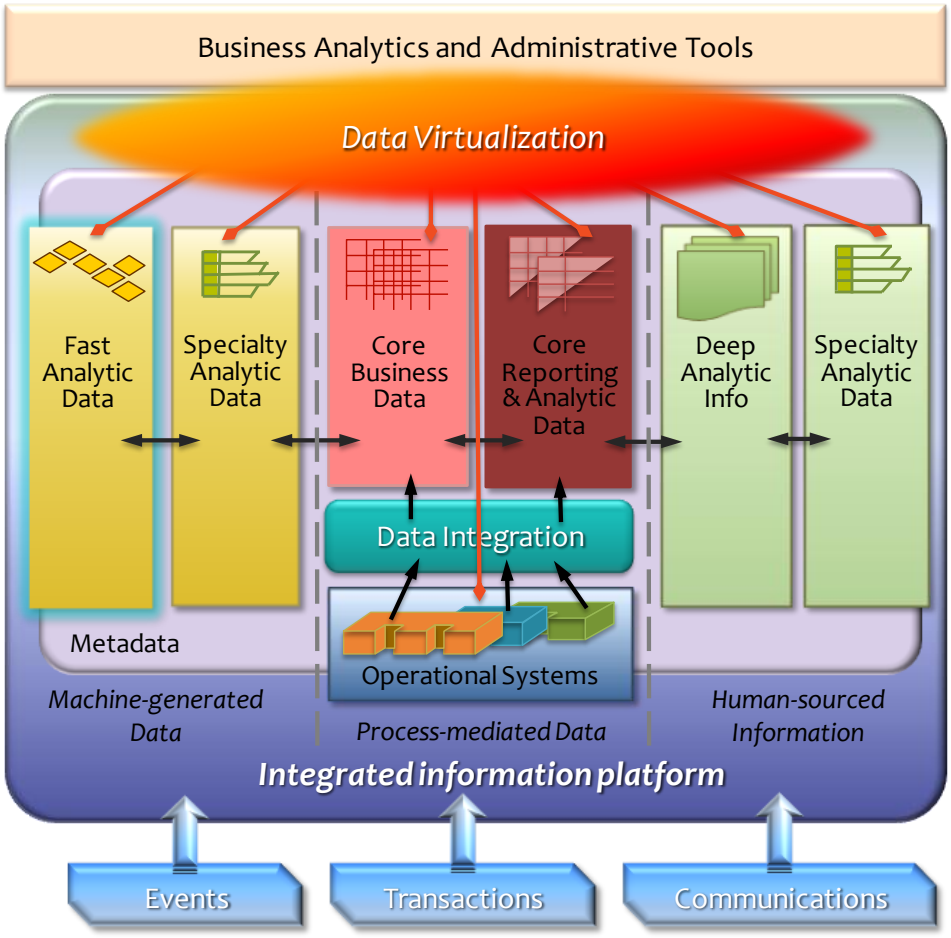
"The eagle may soar; beavers build dams."[8]

If data warehouse developers were animals, they would likely be beavers, industriously taming wild data flows and creating a deep pool of consistent business information. EDWs and associated enterprise data management environments, such as master data management (MDM) systems, are long-proven repositories for well-managed and governed process-mediated data. In contrast, the origins of the big data movement in science and Web companies such as Google and Yahoo! with strong engineering backgrounds, has led to an approach based on Open Source technology and bespoke programming, placing less emphasis on data quality and more on adaptability, scale and speed. Business today needs both sets of attributes; one cannot replace the other.

An integrated platform for all information types, shown in figure 2, must therefore consist of a number of database and analytic technologies, each optimized for a particular type of processing and access, called *pillars* and named according to the business role they support.

1. The first, central ***core business data*** pillar is the consistent, quality-assured data found in EDW and MDM systems. Traditional relational databases, such as IBM DB2, are the base technology. Application-specific reporting and decision support data often stored in EDWs today are excluded.
2. ***Core reporting and analytic data***, the second pillar, covers these latter data types. In terms of technology, this pillar is also ideally a relational database. Data warehouse platforms such as IBM InfoSphere Warehouse, IBM Smart Analytics System and the new IBM PureData System for Operational Analytics, play strongly here. Business needs requiring higher query performance may demand an analytical database system built on massively parallel processing (MPP), columnar databases or other specialized technologies, such the new IBM PureData System for Analytics (powered by Netezza Technology).
3. ***Deep analytic information*** requires highly flexible, large scale processing such as the statistical analysis and text mining often performed in the Hadoop environment.

Figure 2: *The integrated information platform*



4. ***Fast analytic data*** requires such high-speed analytic processing that it must be done on data in-flight, such as with IBM InfoSphere Streams, for example. This data is often generated from multiple sources that need to be continuously analyzed and aggregated with near-zero latency for real-time alerting and decision-making.

5. At the intersection of speed and flexibility, we have ***specialty analytic data***, using specialized processing such as NoSQL, XML, graph and other databases and data stores. It appears twice in the platform because it applies to both machine-generated data and human-sourced information.

Figure 2 shows how these pillars are distributed across the three domains of information and also positions traditional operational systems of record centrally in the platform. The central pillar of the platform is thus closely aligned to the traditional data warehouse architecture, with the important difference that in data marts used for reporting and analysis, data can—and often should—be fed directly from the operational systems.

*The integrated information platform contains **all** the information generated and used by the enterprise.*

Metadata, shown conceptually as a backdrop to all types of information, is central to this new architecture to define information context and enable proper governance. In the process-mediated and machine-generated domains, metadata is explicitly and usually separately stored; in the human-sourced domain it is more likely to be implicit in the information itself. This demands new approaches to modeling, discovering and visualizing both internal and external sources of data and their interrelationships—as seen in IBM Vivisimo's tooling for information optimization—within the platform.

In terms of functionality required, we can see central roles for data integration, which moves, copies, cleanses and conditions data within the platform (black arrows); and data virtualization (orange linkages). Metadata, of course, also plays a key role in both sets of function. Data integration, also known as ETL (extract, transform and load), is familiar from data warehousing and performs the same role in the integrated information platform.

Data virtualization, on the other hand, is anathema to some data warehouse purists. However, unlike the long-established EDW architecture, which insisted that all data flows through a single, physically-instantiated store, the integrated information platform is a set of related stores—logically unified via the core business data and metadata. Data virtualization provides users and applications with access to data stored in disparate technologies and different locations through a semantic layer, offering a business-oriented view of the information, hiding the technical complexity of accessing it and enabling real-time joining of results from multiple sources.

Business analytics and administrative tools include all of the function you might expect, including exploration, visualization and discovery, as well as application development, systems management and so on. Business analytics covers both big data usage as well as traditional BI functionality.

The integrated information platform is the virtual unification of big data and traditional business information.

At its fullest extent, the integrated information platform contains all the information generated and used by the enterprise. This information ultimately comes from the interactions of the business with machines and people, both internal and external, and with other organizations. This is shown at the bottom of figure 2. Events are typically recorded by sensors and machines. Communications are all the interactions that occur between people. And transactions are the subset of interactions that are of financial importance to the business. Transactions are of critical importance to any business; that's why they were the original area for computerization and why operational systems undertake extensive quality assurance work before they are accepted. They are also the main source of core business data. Events and communications require less quality assurance and can thus be loaded directly into the systems that use and analyze them.

Implementing an integrated information platform

How do you eat an elephant? In small pieces…

Like big data, if information in all its forms were an animal, it would also be an elephant… actually, a herd of them. As we've seen, reining in that herd requires an integrated platform spanning all types of information. Although the full vision and scope of this platform is extensive, it already exists in part or, more often, in multiple parts. It is, in fact, a work in progress that began in many organizations with their first EDW, probably as far back as the 1990s, when they started to create information to be used across the organization. One of the primary goals of the original data warehouse architecture[9] was consistency, which is the foundation of enterprise-wide information use—whether of big data or small. Most of the methods and techniques used in building an EDW apply to big data, as do many of the technologies. The important point is to resist the belief that the new technologies are so different that they change everything. They do not.

Starting from your existing systems and, in particular, comprehensive EDWs, you can begin to build the integrated information platform needed to deliver real business value from big data. And you can start right now to take early advantage of some of the key benefits of this platform:

- Reuse of existing data and environments where appropriate
- Agility to introduce new technology as required
- Consistency of information meaning and use across environments
- Improved time to value and return on existing technology investment

To start now with an integrated platform for big data, look to the existing infrastructure and organizations for data management—especially the enterprise data warehouse.

If you are looking to obtain business value from human-sourced information from the Web, such as social media, or from internal sources, such as call center logs or extensive text archives, building a sandbox environment in Hadoop can make good sense. From a technology viewpoint, it is vital that this new environment is linked as closely as possible with your existing business intelligence (BI) system to allow two-way transfer of information—certified core business data about customers or products, for example, into the Hadoop environment to frame the analysis; and summarized data from analytic tasks into the BI system to drive reporting and decision-making processes.

From an organizational point of view, much focus has been placed on data scientists and their scarcity in market. Data scientists are experts in solving complex data problems using a combination of skills such as data collection and cleansing, statistical analysis, visualization and deep domain knowledge. One place that is sometimes overlooked as a source of data scientists is among the power users of BI systems and spreadsheets within the business departments that currently use data most extensively. Those in the marketing department, who have a line-of-business mindset and are used to analyzing lots of data and making sense of it, are often a good fit. They may need training in more advanced statistical or programming methods, but they have domain knowledge and the right mindset. If bringing new data scientists on board, it is vital to ensure close involvement of the BI team in the new environment to guarantee that these new skills will be well integrated into existing teams. A good approach is to build a team of 2-3 members—one from the business, one who understands analytics and likes to play detective, and an IT expert from the BI team, who can access data from the EDW and integrate it into the new big data technologies.

If your business focus is directed towards new insights from or processes around machine or sensor data, your choices are wide. You might begin with an analytic database, such as the new IBM PureData System for Analytics, to store and explore this data or if you have more operational analytical needs, the new IBM PureData System for Operational Analytics. You could also use Hadoop and IBM InfoSphere BigInsights if the volumes are particularly large or the structures very variable. If speed of processing is the requirement, a streaming solution, such as IBM InfoSphere Streams could be the answer. In any case, the technological and organizational imperatives are the same as those mentioned above—close integration with the BI environment and team.

As you address additional business needs and add new functionality, one of the key benefits of a platform approach will quickly become apparent—reuse of infrastructure and data resources. The same data integration and metadata will be used across the different parts of the platform. Quality assurance work done in one area will improve data quality throughout. Business users will have broader access to different types of data—if their job demands—through a common set of tools with more consistent usage and a greater level of contextual meaning.

In some senses, big data poses many of the same types of problems for data management as have spreadsheets. Mention spreadsheets to most BI or data governance groups and they run for cover! As Wayne Eckerson lamented: *"Spreadsheets run amuck in most organizations. They proliferate like poisonous vines, slowly strangling [them]..."*[10] Big data, as often implemented today, is similarly uncontrolled, unmanaged and centered around individual data scientists and their tools and data sets. Implementing an integrated platform is an important step in limiting a similar proliferation. Together with close integration with the existing BI organization, this approach can make big data into a powerful tool for innovation and process evolution, rather than weapons of mass value destruction.

But perhaps the single, most important step towards implementation is to obtain business buy-in and sponsorship at the highest level. This, of course, is old news for BI developers. But, beware! The background of some big data vendors and advisors is more from the programming, Open Source and Web development communities, where this enterprise-level, business sponsorship is seldom the norm. Associating your big data initiatives with *previously successful* EDW and BI initiatives is likely to be the best approach to gain credence with the business. With the substantial and often highly visible business benefits offered by big data, executive buy-in at the highest levels can be easier and quicker to obtain than in the case of EDW initiatives. Such enthusiasm can and must be used to support the implementation of an integrated information platform. And learning from prior experience, in a staged, incremental approach that delivers business benefit at each step.

As in the case of BI, business buy-in and executive sponsorship are the most important success factors for big data implementation.

Conclusions

Big data offers probably the most significant, big game-changing business opportunities that have been seen since the emergence of Web-based e-business in the late 1990s. Of course, big data has been over-hyped in much the same way as was e-business. However, we have reached a turning point. A more realistic attitude has emerged as traditional information management vendors have become more involved in the market and the focus has moved from Internet startup businesses to well-established, mainstream enterprises. It is increasingly clear that big data is best implemented as part of long-standing, overall information management processes and focused on business outcomes. Why? Because big data, whatever its extreme volume, velocity or variety, is simply more business data that must be appropriately managed like all other business data and integrated with existing sources. Big data on its own can offer *ah-ha* insights, but it can only reliably deliver long-term business advantage when fully integrated with traditional data management and governance processes.

The moment has arrived when big data moves from bleeding- to leading-edge. More mainstream businesses are taking advantage of the opportunities offered by big data to reinvent key decision-making and operational processes. Central to this evolution is the creation of a big data platform supporting many types of big and small data in an integrated, enterprise-grade environment with business analytics that can operate directly on data in its native format, and as close to the data sources as possible. The key business advantages of such an integrated platform are:

1. ***Provide predictive insights to future outcomes*** by grounding social media and customer behavior analysis in real, quality customer data that the enterprise has long collected for daily use
2. ***Drive real-time operational decisions*** with faster insights in a broader context from machines and sensors in the external environment, used in conjunction with traditional transactional data
3. ***Reinvent business processes*** for faster, innovative action and game-changing business models by closing the loop between informational and operational activities

With such significant business benefits at stake, IT must—and, indeed, can—make a rapid and incremental start by building from the existing data management infrastructure. In many cases, the starting point is the current data warehouse and BI environment. Examples include: introducing Hadoop to pre-process and analyze existing content, such as call center records; adding stream computing to bring real-time data into the data warehouse; and revamping an existing data warehouse to feed analytic databases directly from sensor data sources. There exist numerous entry points to this new platform that require only relatively small investments in time, effort and money to deliver early and real business results and allow IT to build from there.

To take advantage of these very real opportunities, collaboration between business and IT is vital for an immediate start to plan and deploy a comprehensive but incremental big data strategy. Starting small with agile project methods will deliver early business value and bring analytics and data scientists into the mainstream. As big data tooling has matured and becomes more closely integrated with existing data management platforms, this is the moment when innovative companies can break from the pack and gain the most immediate and long-lasting competitive lead.

An enterprise integrated information platform is the first step towards taming the big data beasts and deriving real, long-lasting business benefit from the current zoo.

Dr. Barry Devlin is among the foremost authorities on business insight and one of the founders of data warehousing, having published the first architectural paper on the topic in 1988. With over 30 years of IT experience, including 20 years with IBM as a Distinguished Engineer, he is a widely respected analyst, consultant, lecturer and author of the seminal book, "Data Warehouse—from Architecture to Implementation" and numerous White Papers.

Barry is founder and principal of 9sight Consulting. He specializes in the human, organizational and IT implications of deep business insight solutions that combine operational, informational and collaborative environments. A regular contributor to BeyeNETWORK, Focus, SmartDataCollective and TDWI, Barry is based in Cape Town, South Africa and operates worldwide.

Picture credits:
African elephant: Barry Devlin
Blind men: C. M. Stebbins & M. H. Coolidge, "*Golden Treasury Readers: Primer*", American Book Co. (New York), 1909 [Wikipedia.com]
Eagle: www.123rf.com/photo_5236964_american-bald-eagle-in-flight-blue-sky-on-background.html [LoonChild / 123RF.com]
Beavers: Willem Janszoon Blaeu: "*Nova Belgica et Anglia Nova*" (Detail), 1635 [Wikipedia.com]
Origami elephants: Katherine Devlin
Chauvet cave painting: HTO [Wikipedia.com]

---

[1] "*Data, data everywhere – A special report on managing information*", The Economist, February 2010

[2] "*Outperforming in a data-rich, hyper-connected world*", IBM Center for Applied Insights, March 2012, http://bit.ly/MKxHhe

[3] "*Big data: The next frontier of innovation, competition and productivity*", McKinsey Global Institute, May 2011

[4] "*Gartner Reveals Top Predictions for IT Organizations and Users for 2012 and Beyond*", Gartner, December 2011, http://bit.ly/S2mvgW

[5] From the poem "*The Blind Men and the Elephant*" by John Godfrey Saxe (1816-1887)

[6] "*Expanding Digital Universe*", International Data Corporation (IDC), 2007-2011, http://bit.ly/IDC_Digital_Universe

[7] Euripides, Greek playwright (c. 480-406 BC)

[8] Joseph S. Nye, Jr. (1937-)

[9] Devlin, B. A. and Murphy, P. T., "*An architecture for a business and information system*", IBM Systems Journal, Volume 27, Number 1, Page 60 (1988) http://bit.ly/EBIS1988

[10] Eckerson, W., "The Rise and Fall of Spreadmarts", DM Review 2003