

COLLABORATIVE WHITE PAPER SERIES

Making Sense of Big Data

A Collaborative Point of View



COLLABORATIVE WHITE PAPER SERIES

Making Sense of Big Data

A Collaborative Point of View

Primitive organisms dominated life on Earth for 3.9 billion years and then in a geologic wink, life evolved rapidly into a vast array of complex organisms. Over the course of geologic time rich deposits of organic matter was buried under the earth's strata and subject to various forces that in time and with the development of new technologies enabled this matter to be used as a source of fuel for industry. Improvements in technology over time have made it possible for industry to locate and retrieve more and more of these fuel sources; some of which until recent time were viewed unreachable or even unusable.

Similarly, within "computing time" more data has been created and deposited within the layers of corporations and mined to drive business growth. At this particular stage, an extremely large quantity of data is being deposited as a result in the explosion of social media and digitization of existing business processes. This transition has given birth one of the latest buzzwords: "Big Data".

As with every new buzz in the technology arena, much of it is driven by the vast marketing engine of software and services companies. However, this unrefined resource has presented a new source of opportunity if it can be recovered and used. While this is true with Big Data, the similarities between the Big Data "buzz" and energy are unmistakably. The growth in data coincides with the development of new technologies to enable the recovery, storage and to a lesser extent refinement of that data so it becomes an asset that fuels business growth. Therefore, Collaborative views Big Data as a paradigm shift in the Information Management industry, allowing valuable analytic gains from previously underutilized data.

This Collaborative Point of View will present you with the many dimensions of Big Data to help you identify areas in your business where you can put its power to work, or to the contrary, decide that it does not apply to the challenges you are facing collecting and analyzing data. You will read that many of the concepts of Big Data have been around for a long time and are now being viewed through a new lens, due to the powers of the enabling technologies. However, many of the concepts are new to most industries, such as gaining insight from large volumes of unstructured data. Some concepts use the same term but are applied differently, such as Data Governance. If the reader is to take one point from this article, it is that Big Data is not one thing, and it does not have magical powers to provide advanced business analytics on its own. Big Data is not a technology. It is, however, a shift in the thinking on how to gain insight from data with increasing volumes and varying formats. It is enabled by exciting new technologies and evolutionary applications of mature technologies.

Big Data is not one thing, and it does not have magical powers to provide advanced business analytics on its own. Big Data is not a technology. It is, however, a shift in the thinking on how to gain insight from data with increasing volumes and varying formats.

For organizations trying to decode the DNA of this new species of data, formulating a reasonable strategy for Big Data can be intimidating. It's no accident that the polarizing era of "Big" (e.g. Big Oil, Big Government) coincides with the emergence of Big Data. Despite its imposing nature, organizations can make sense of Big Data, but only if they take a step back to think about:

- The etymology of the term "Big Data"
- What makes data "big"
- How Big Data differs from conventional data
- How Big Data could yield business benefit
- What is the current state of Big Data technology and vendors
- What is the correct path to Big Data adoption

Etymology: A short guide to buzzword bingo

When examining the fossil record of Big Data, there appear to be several events that will help explain its origin. The term "Big Data" was popularized, if not invented, by Roger Magoulas of O'Reilly in 2005; however, mentions of the term date back to the early 2000's. In 2004 Oracle wrote a paper on the concepts of MapReduce¹, a software framework to perform distributed computing on large data sets. This concept gained popular momentum with the Apache open source project Hadoop, whose first beta release was available in 2007.

The term Big Data and Hadoop quickly became linked. Popular interest in NoSQL was reinvigorated in 2009 and attached to the term Big Data. NoSQL has been around since the early 90's; these databases store data in non-traditional ways and introduce new query languages for access. Hadoop is often categorized as a NoSQL implementation. Around this same time the major BI and ETL vendors began hitching their carts to the Big Data wagon.

Respected analyst firms such as Gartner and The McKinsey Global Institute began publishing thoughtful papers on Big Data. McKinsey's paper "Big data: The next frontier for innovation, competition and productivity", published in May of 2001², is a must read for anyone serious about gaining a deep understanding of the business application of Big Data.

Next comes HortonWorks, whose focus is accelerating the adoption of Hadoop. Existing database appliances, such as Netezza and Teradata, take on the Big Data moniker. Traditional relational databases, such as Oracle and SQL Server, announce Big Data implementations, with some including Hadoop back-ends. Terms such as "Big Data Analytics" and "Extreme Information Management" are attached to the Big Data moniker.

On and on and on the buzz continues. As each thoughtful software and services company defines its value proposition as it applies to Big Data, the waters become increasingly muddled. This POV tries to provide a level of clarity to the uninitiated and to the casual observer alike.

Big Data is...

The nature of data has changed. In short, there is a rapidly expanding universe of data that companies could absorb. The latest IDC Digital Universal Study reveals an explosion of stored information: more than 1.8 zettabytes—or 1.9 billion terabytes—of information was created and stored in 2011 alone. To put this number in perspective, this means that 61.3 terabytes of information was produced *per second* over the course of a year. Experts predict that by 2020, there will be a 42x growth in data from 2009³.

One way to frame the primary concepts of Big Data is to look at Big Interactions, Big Transactions, Big Processing and Big Analytics. In short, Big Analytics are enabled by the ability to processes large volumes of transaction and interaction data that vary in format and frequency of change.

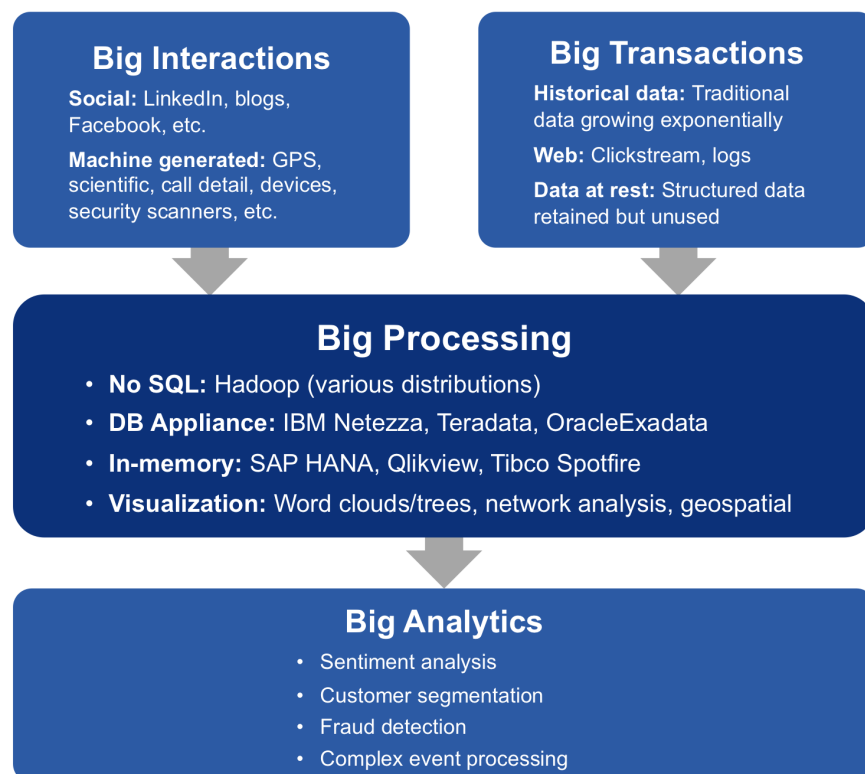
Part of the digital explosion is caused by the natural growth rate of structured data from new channels (e.g., machine instrumentation, RFIDs, location awareness, mobile devices, aka the "Internet of Things"), and new markets (global growth). There are other drivers of data growth that tend to be industry specific. Regulatory forces (Dodd-Frank, EU's transformative Markets in Financial Instruments Directive), algorithmic trading, and market fragmentation are the key factors behind the data growth in capital markets⁴. However, the majority of this

growth is coming in the form of unstructured or loosely-structured data such as social media, audio, personal blogs, or SMS. The rapid ascension of both new structured and new unstructured data is driving this staggering growth in raw information, and will continue to do so in the foreseeable future.

One way to frame the primary concepts of Big Data is to look at *Big Interactions, Big Transactions, Big Processing* and *Big Analytics*. In short, Big Analytics are enabled by the ability to processes large volumes of transaction and interaction data that vary in format and frequency of change. The popular three V's of Big Data is a good place to start: Volume, Velocity and Variety.

- **Volume:** The Big in Big Data. It is hard to disagree with the fact the sheer volume of data is exponentially growing. McKinsey's paper on Big Data sums it up well:
 - \$600 to buy a disk that stores all of the world's music
 - 30 billion pieces of content are shared on Facebook every month

- 40% projected growth in global data generated each year
 - 235 terabytes of data collected by the Library of Congress by April 2011
 - 15 out of 17 sectors in the United States have more data than the Library of Congress
 - \$300 billion potential annual value to US healthcare
 - 140,000-190,000 more deep analytic talent positions are needed
 - 1.5 million more data-savvy managers are needed to take full advantage of Big Data in the U.S.
- **Velocity:** The Big in Big Data is not the only challenge. A form of data that is described later in this article is machine-generated data, essentially data generated by machines talking to other machines. Think of the sheer velocity that is accomplished from the streaming data produced in the various devices and sensors we come into contact with every day. While this data is not new, tapping it in aggregate to gain analytic insight is enabled by the advent of Big Data. Velocity is not only generated with machine data; the ability to take real-time transactional data and make analytic decisions is also enabled by Big Data.
 - **Variety:** Another non-Big in Big Data is the ability to process data generated in a variety of formats. Variety is what makes Big Data pertinent to many businesses where volume is less of a challenge. The variety includes the traditional structured data we have been working with for decades, semi-structured data in formats like XML, and unstructured data such as Tweets, Facebook, blog posts, audio, video or images. Enabling analytics on the unstructured data generated as part of the social media explosion is the sexy part of Big Data – for example, the ability to decode a person's sentiment toward your product. Equally profitable is making sense of less sexy forms of unstructured data. For example, insurance companies regularly extract facts from text gathered in claims processing, then use that data to for risk management and fraud detection.



Big Data is different because...

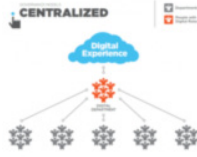
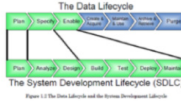
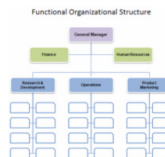
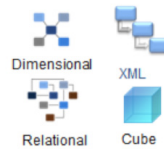




Big Data doesn't fit neatly into the conventional data boxes (e.g. relational databases and data models, or conventional governance models), and as a result it is causing organizations anxiety as they try to separate information gems from the mountains of coal. The conventional models of structured, engineered data do not sufficiently reflect the realities of Big Data. The key to leveraging Big Data is to understand these differences before expediting its use.

The most significant difference is that data is mostly governed in a centralized manner, but Big Data is self-governing. These contrasting governance models explain the fact that conventional data evolves more slowly until it is retired, and is typically defined within the context of a SDLC for a business application. Big Data evolves rapidly in the context of topical "Big Interactions" until the subjects become extinct, or morph into new topics. The iterative Big Data lifecycle can be represented in 5 phases: Prospect, Recovery, Storage, Refining, and Application. The sourcing model is also very different between data and its bigger cousin. Data is mostly created and used by a workforce of trained professional users. Big Data is created by a rapidly expanding universe of machines or users of highly varying expertise and savvy. Consequently, the composition of traditional data will vary drastically from its larger cousin. The composition for data serve a specific purpose, and must be more durable and structured, whereas Big Data will cover many topics, but not all topics will yield useful information for the business, and thus they will be sparse in relevancy and structure.

The conventional models of structured, engineered data do not sufficiently reflect the realities of Big Data. The key to leveraging Big Data is to understand these differences before expediting its use.

The following illustrates the contrasting governance, lifecycle, sourcing, and composition profiles that separate the two species of data.

Figure 1: The varying profiles of Data and Big Data

	Governance model	Lifecycle model	Sourcing model	Composition
Data	Organized central governance 	Slowly evolving to retirement in a SDLC context ⁵ 	Trained professional hierarchical workforce 	Purposeful, concentrated, structured 
Big Data	Decentralized self-governance ⁶ 	Rapidly evolving to extinction in a topical context 	Crowd-sourcing 	Broad, sparsely concentrated, unstructured 

The value of Big Data is...

In the world of Big Data, more is more, i.e. more data can yield increasing value to the business. Exploring the universe of Big Data can be a rewarding proposition, provided there is sufficient business involvement and focus. Focus is critical because the pervasive nature of Big Data could lead an IT operation down many dead ends, without a cohesive business roadmap to indicate what type of Big Data could serve the greater good.

Given the potential of Big Data, it is easy to forget that sometimes from a data perspective, less is more. The right amount of high quality data provides more insight than volumes of poor quality data. Big Data will not improve data quality; if anything, the velocity of suspect data will increase. Thus, the potential for more data increases the need to make sure the data is reliable. The risk is that if organizations expose users to Big Data without the proper processes in place, bad information will leak into the hands of users, resulting in poor tactical or strategic decisions.

Focus is critical because the pervasive nature of Big Data could lead an IT operation down many dead ends, without a cohesive business roadmap to indicate what type of Big Data could serve the greater good.

Prospecting Big Data will expose users to toxic data. Not all toxic data may be viewed as toxic in the future. Consider that in the 1890's, an unvalued byproduct of the refining process, gasoline, was openly dumped into the Cleveland River much to the dismay of riverboats that cleaned their steam boilers at night near the gasoline dumping grounds. Thus companies may need to reassess their threshold of data quality to maximize the potential of this unrefined asset. During the refinement process, companies can align Big Data to trusted master data domains to establish the linkage and relationships from unstructured data to the master data so you can effectively analyze the data. Thus, a strong multi-domain MDM program and platform can help make sense of Big Data.

Big Data will not improve data quality; if anything, the velocity of suspect data will increase.

If organizations can channel sources of Big Data into information that is qualified and available for use, it could provide users with new ways of looking at customers, products, vendors, competitors, or themselves. Big Data augments existing analytic subject areas like customer segmentation and fraud detection, and spawns new analytic domains such as those listed below.

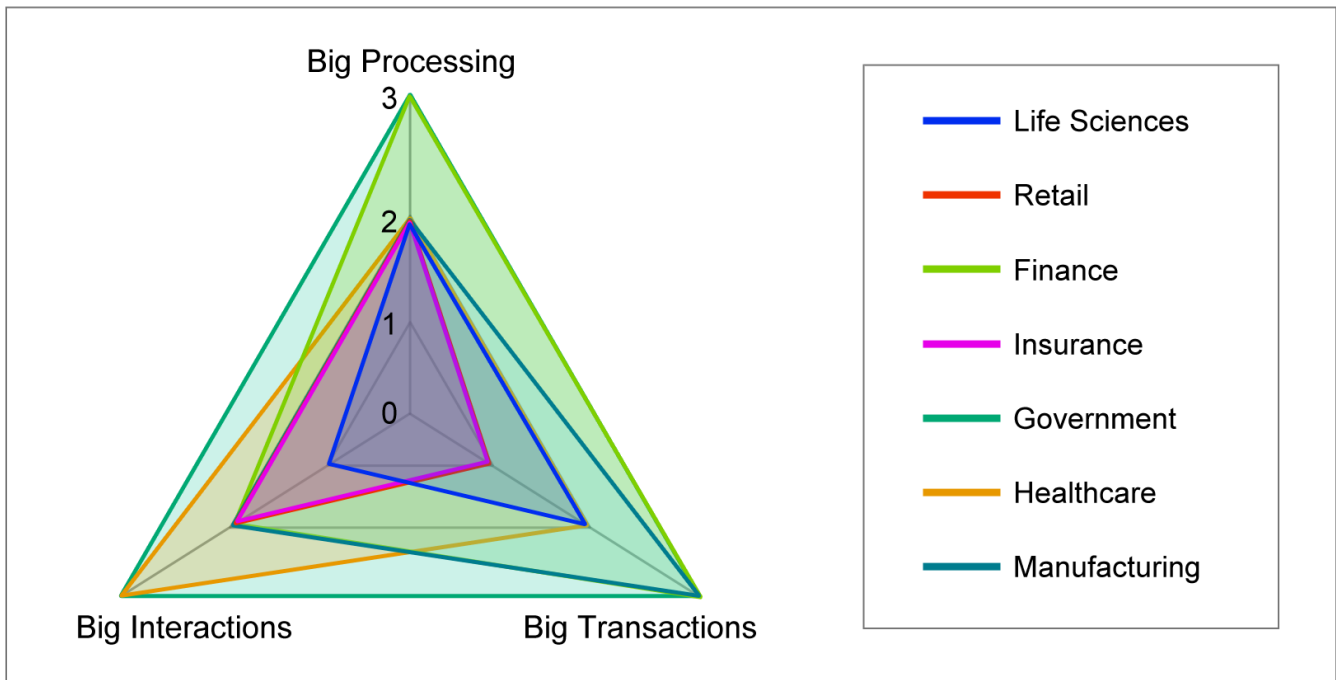
Big Data analytics	Description	Value proposition	Use case
Sentiment analysis	Mining semi-structured or unstructured data text into discernible and measurable facts to a specific topic or event. Heavily dependent on Big Interactions and data visualization (e.g. word clouds & theme maps).	Improved visibility into market attitude on products, services, promotional activities or an organization as a whole.	Social media responses to a particular promotional campaign can provide enhanced visibility into marketing effectiveness, influencing new offers targeted to respondents, or perhaps suggesting new features or enhancements to product development.
Customer segmentation	Big Interactions open new doors for segmenting customers, such as applying an "influencer" score, that applies a rating to active participants on blogs and social media sites that are heavily followed.	Allows companies to identify a specific audience that dictates trends and attitudes, and helps organizations align the segment to existing customers and product offerings.	Marketers and product managers that can track influential trends and tailor products and services that favor the high-end of influencers can improve the overall quality and popularity of items brought to market.

Big Data analytics	Description	Value proposition	Use case
Fraud detection	Exposing the sinister side of Big Interactions (emails or machine data), Big Processing (e.g. statistical modeling), and mining Big Transactions to detect fraud.	Lowers risk, allowing organizations to preempt schemes intended to defraud customers, vendors, investors, or other key stakeholders.	Pillars of e-Payments mine Big Data to detect fraud. PayPal and Amazon align IP addresses to suspect transactions to identify potentially fraudulent transactions. ⁷
Financial Risk Modeling ⁴	Regulators' increased interest in vetting firms' risk-based capital calculations, stress test results, Value at Risk (VaR) computations and other metrics will boost demand for data to feed these exercises, as will regulators' growing scrutiny of the inputs to firms' Level 3 asset valuations.	In-database statistics and MPP allows companies to increase the complexity of financial risk models, and deliver results in a fraction of the time.	Stochastic Monte Carlo VaR calculations, which at one time took firms' supercomputers all night to run, are now being cranked out much more frequently, thanks to increases in off-the-shelf computer power and a steadily growing library of high-performance software tools
Logistical / manufacturing efficiency	Machine-generated Big Interactions and Big Processing coupled with advanced data visualization.	Lowers operational costs, improves quality, and increases speed-to-market.	Tracking GPS data (i.e. data at-rest) for historical routes can identify the most efficient delivery schedule and movements for transportation fleets.
Complex event processing	Social media in particular provides organizations with a source of information that is incredibly rapid and intense in response to current events.	Lowers risk and improves responsiveness to worst-case scenarios (e.g. natural disasters, financial downturns).	<p>Recently tweets about a rare 5.8 magnitude earthquake in West Virginia outpaced the seismometers of the U.S. Geological survey, and the first tweets of the quake reached New York 40 seconds ahead of the quake's initial shock waves.⁸</p> <p>Recognizing the voracious ability of social media to stay on top of rapidly moving trends and events, the US Geological survey is now experimenting with Twitter as an effective way to monitor earthquakes.⁹</p>

Certain organizations stand to benefit more than others from the use of Big Data. For instance, a pharmaceutical manufacturer relies heavily on high-volume, structured third-party data for commercial data analysis. The data volumes, while high, can typically be handled by most RDBMS platforms or data warehouse appliances from a Big Processing and Big Transaction perspective. Thus, Big Data may not make sense to the commercial operations of a pharmaceutical company. In that same company, the clinical and market research operations may rely heavily on the use of advanced statistical analysis of semi-structured data. In that sense, the Big Processing and Big Interaction aspects of Big Data serve the pharmaceutical company well.

At a high level, each industry aligns to the Big Data pyramid in terms of relevancy. Big Data varies in its relevancy to a given industry in terms of Big Interactions, Big Processing, and Big Transactions. The following relevancy index scores major industries along these key dimensions of Big Data.

Figure 2: Big Data relevancy index by industry



Big Interaction scale	Big Transition scale	Big Processing scale
0: No impact	0: <200 petabytes	0: Not applicable
1: Low impact	1: 200-400 petabytes	1: Low applicability
2: Medium impact	2: 400-600 petabytes	2: Medium applicability
3: High impact	3: >600 petabytes	3: High applicability

Source: MGI Big Data Report, p 20 & 22

Source: MGI Big Data Report, p 19

The current state of Big Data technology is...

Emerging technical platforms are providing enhanced capabilities that facilitate the understanding and analysis of Big Data. These technologies are allowing larger data sets to be integrated and analyzed on a massive scale, and enable enhanced visualization and statistical modeling that present the vast amounts of information in understandable terms.

There are a plethora of technologies and vendors that can be labeled as purveyors of Big Data, which makes it very confusing for organizations that hope to capitalize on the business value of Big Data. The current field is crowded and chaotic. A cursory survey of Big Data technology and vendors would resemble the heavy bombardment period of the Universe; a rapid emergence of vendors varying in size coalescing into larger objects. In other words, the list of vendors will shrink as larger vendors purchase smaller vendors. As shown later, some niche Big Data vendors have already been purchased by larger entities.

A cursory survey of Big Data technology and vendors would resemble the heavy bombardment period of the Universe; a rapid emergence of vendors varying in size coalescing into larger objects.

Before diving into the deep end of the Big Data technology pool, let's quickly revisit the relatively rapid evolution of the enabling technologies behind Big Processing, Big Transactions, Big Interactions, and Big Analytics. The confluence of the following milestones has given birth to new methods and deployment models for Big Analytics:

- **NoSQL platforms:** Search engines have invented new methods of processing massive amounts of unstructured data for search. Google patented MapReduce on their BigTable file system, which inspired the Yahoo!-driven open source project for MapReduce on the Hadoop Distributed File System (HDFS). The result is a new way of crunching massive data sets in a fault tolerant grid of commodity hardware, and ushering in the NoSQL movement.
- **Appliances:** The original Big Processing platform – typified by industry-leading MPP and easy-to-maintain hardware / software combinations – has been around for decades. Early pioneers like Teradata were joined by Netezza (now IBM), and appliances from Microsoft, Oracle, and EMC to support highly scalable Big Processing environments.
- **In-memory analysis:** Perhaps the most subtle innovation is one of the most profound and beneficial to users. The ability to compress massive amounts of data into memory provides users with the performance of a cube and unprecedented speed-of-access to detailed granular facts, usually completely disrupting the architecture and design of an existing data warehouse.
- **The Internet of Things:** Machine-generated data from networked devices, instrumentation, RFIDs, etc., have produced a new source for Big Interactions that operate on a non-stop basis, leading to an untapped source of raw data that drives Big Analytics.
- **Social Networking explodes:** Blogs and social media have contributed to the explosion of Big Interactions that are an important source to Big Data analytics.
- **SaaS model grows:** Salesforce.com proved that critical business applications did not have to run on-premise, resulting in a new software deployment and pricing model. Big Data SaaS companies like 1010data and traditional giants like IBM are demonstrating that the SaaS method of multi-tenant, off-premise Big Data deployments are a viable alternative to heavy up-front on-premise investments.

The following table lists several key capabilities of Big Data platforms, namely:

- *In-Memory:* Storing massive amounts of data in high-speed memory
- *NoSQL (platform):* A catchall for a commercial non-relational database (e.g. Hadoop)
- *NoSQL-Aware:* The ability to connect into NoSQL platforms for data integration or BI
- *Appliance:* A convenient, MPP-class server, storage, and software combination
- *Big Data Integration:* The ability to integrate and load from / to Big Data sources
- *Visualization:* Applying enhanced visualization techniques such as geospatial maps, word clouds, and theme maps that help users absorb semi-structured or unstructured data
- *Big Data Analytics:* Predefined solutions that provide analysis for domains specific to Big Data sources, such as social media monitoring
- *In-database statistics:* Support for advanced in-database statistical programming based on proprietary languages (e.g. S+) or open standards (e.g. R)
- *Big Data SaaS:* The ability to support a given Big Data capability in a secure, off-premise deployment

This table should illustrate a key trend, SaaS software models optimized for the cloud have given rise to the biggest growth areas for Big Data – social networks – that allow for the rapid creation and sharing of information. As shown below, this SaaS-model is also emerging as a viable platform for Big Data.

Figure 3: The crowded field of Big Data players

● Current ◐ Partial ○ Planned

	Description ¹⁰	In-Memory	NoSQL (platform)	NoSQL-aware	Appliance	Integration	Visualization	Big Data Analytics	Big Data SaaS	In-database statistics
1010data	Offers a web-based service and the underlying software that makes it easy to acquire, organize, manage, and analyze large volumes of complex, interrelated data								●	
Amazon Web Services	Provides companies with an infrastructure web services platform in the cloud. For cloud-based Big Data processing, AWS offers DynamoDB, a proprietary NoSQL database, its Elastic MapReduce Service, and its S3 storage service		●						●	
Clarabridge	Provider of text analytics software for customer experience management						●	●	●	
Cloudera	Develops and distributes commercial Hadoop		●	●						● (R)
EMC	Offers Greenplum that supports multi-terabyte data warehousing demands		●	●	●					● (R, SAS)
Google (PostRank)	Monitors and collects social engagement events with online content in real-time across the web							●	●	
Horton-works	Makes Hadoop more robust and easier to install, manage and use. The company also provides support and training for Apache Hadoop		●							
IBM	A leader in BI, DB, and data integration	●	●	●	●	●	●	●	●	● (R, SPSS, SAS)
Informatica	A leader in data integration, MDM, and data quality software			●		●			●	
Klout	Identifies influencers and provides tools for influencers to monitor their influence, a new way of segmenting customers								●	
Kognitio	Offers a MPP database that is optimized for data analytics, data warehousing and business intelligence	●			●			●	●	
Microsoft	A leader in BI, DB, and data integration	●	○ (2012 Q2)	○	●	○ (2012 Q2)	◐		●	

	Description ¹⁰	In-Memory	NoSQL (platform)	NoSQL-aware	Appliance	Integration	Visualization	Big Data Analytics	Big Data SaaS	In-database statistics
Micro-Strategy	A leader in BI	●	●	○			●		●	
Oracle	A leader in BI, DB, data integration, and ERP	●	●	●	●	●	●	●	●	● (R)
Pentaho	Business analytics suite includes capabilities for data access, integration, discovery, analysis, reporting, visualization and mining		●	●						
Qlikview	QlikView is a highly interactive business intelligence software solution based on in-memory associative search technology which allows users to use data to make important decisions	●								
Quantivo	Leading provider of on-demand big-data analytics solutions		●	●				●	●	
Salesforce (Radian6)	Provides the social media monitoring platform with Integrated social media, CRM, and web analytics also help companies define their social media ROI.							●	●	
SAP	A leader in BI, data integration, and ERP	●						●	●	● (R)
SAS	Provides business analytics software and services, and is the largest independent vendor in the business intelligence market	●					●	●	●	● (SAS, R)
Splunk	A leading provider of operational intelligence software used to monitor, report and analyze real-time machine data as well as terabytes of historical data—located on-premise or in the cloud			●				●	●	
Tableau	Provides easy-to-use software applications for fast analytics and visualization						●	●		
Teradata	Leader in data warehousing and analytic technologies		●	●	●	●		●		● (R, SAS)
TIBCO	A leading vendor of ESB software and offers Spotfire for in-memory BI and data visualization	●					●		●	● (S+)

From the table above, we see the usual suspects of vendors and new vendors likely to be acquired in the very near future are packaging some form of emerging Big Data capability – such as NoSQL support, in-memory analysis, or data visualization – into existing offerings. It is important to understand particularly from an IT perspective that none of these platforms at this point in time are a one-size-fits-all solution for Big Data. IT should be mindful of the limitations of nascent Big Data technologies. NoSQL platforms like Hadoop are so heavily hyped at the moment that it is difficult to say whether its current experimental use will evolve into a permanent presence for IT. For instance, if low-latency analysis of machine sensor data is a requirement, Hadoop – specifically MapReduce which operates in batch – would not provide the required capability.¹¹

There is no way of knowing whether companies need NoSQL support, in-memory analysis, or a mix of the two without considering how these platforms improve the ability to serve the business community and how they fit in an existing technology footprint. Also, traditional DW and BI operations could be disrupted without ensuring support for standard query operations like multi-pass SQL. Investing in a Big Data platform may be overkill if an existing data warehouse can support the long-term projected data volumes and data sets, or the business is not relying on Big Data sources. Ultimately, tools alone cannot answer Big Data questions, and the skills to operate these tools are sparse. Throwing technology at a problem without the necessary business justification, or the supporting people and processes in place that can make effective use of the tools, would ground any Big Data initiative.

Ultimately, tools alone cannot answer Big Data questions, and the skills to operate these tools are sparse. Throwing technology at a problem without the necessary business justification, or the supporting people and processes in place that can make effective use of the tools, would ground any Big Data initiative.

Big Data adoption will follow several paths...

In the process of translating Big Data hype, organizations should recognize there are three paths towards Big Data adoption. Two paths are the extremes of the adoption scale. On one end of the spectrum, there is a non-adoption path. On the other end, there is a rapid adoption path. In between, there will be a slow-and-steady path towards Big Data adoption. Organizations will follow one of these paths based on four factors: a compelling business case, supporting data governance, maturing vendor capabilities, and Big Data skills available in the workforce.

As mentioned earlier, Big Data does not offer the same potential for all organizations and departments. For those that don't stand to benefit from Big Data, the business case does not exist and non-adoption makes the most sense. For organizations that could benefit from Big Data, there are two paths to explore: rapid adoption and the "slow and steady" path.

The rapid adoption path (i.e. where companies dive head-first into Big Data investments) makes sense in two scenarios. The first scenario involves organizations that rely on Big Data technologies for core business operations. Companies like Google and Facebook offer products and services enabled by the Big Data technologies listed above. The second and more likely scenario involves organizations that want to leverage Big Data for advanced analytics that have been defined in clear business terms and whose internal data governance capabilities and resource skillsets are strong enough to support immediate Big Data initiatives. The rapid adoption path requires a critical mass of business justification, data governance maturity, and Big Data IT know-how to justify the launch of Big Data projects.

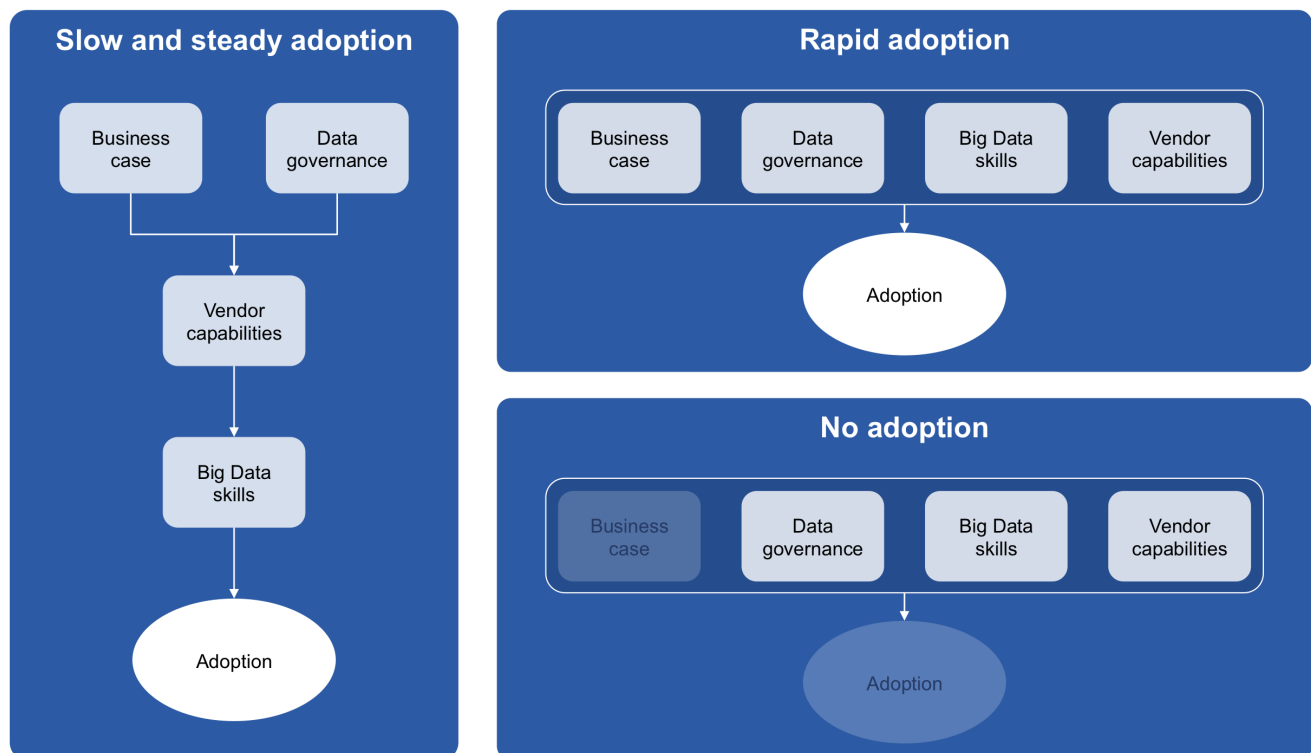
In between these two extremes lies the path more likely to be embraced by organizations on the way towards Big Data adoption. There are several reasons why the "slow-and-steady" path will be the predominant adoption strategy. First and foremost, organizations need time to state a clear business case for Big Data. Until this coherent vision is formulated at a sufficient level, organizations cannot intelligently embrace a non-adoption or rapid adoption path. In addition, organizations must define data governance task forces and allow these nascent organizations to mature before thrusting them into the Big Data spotlight. Another reason organizations will lengthen the adoption

cycle is that vendors have not fully rationalized their Big Data offerings into a fully rationalized solution. Organizations will therefore stay away from the bleeding edge of Big Data technologies until vendors have established a proven track record in this emerging space. Finally, there is a gap between the technology and professionals with the technical skills and experience required to implement and administer Big Data technologies. Unless companies invest in training and can tolerate a trial-and-error period as less experienced trainees perfect their techniques, the growth of a skilled Big Data workforce will lag behind the Big Data adoption rate.

The following graphic represents the Big Data adoption paths.

Big Data does not offer the same potential for all organizations and departments. For those that don't stand to benefit from Big Data, the business case does not exist and non-adoption makes the most sense. For organizations that could benefit from Big Data, there are two paths to explore: rapid adoption and the "slow and steady" path.

Figure 4: Big Data adoption paths



Focusing on data governance slow-and-steady adoption path, organizations need to rethink data governance guidelines and processes that will allow for the proper use of Big Data. Similar to the role of data governance in MDM, where the absence of strong governance can derail the well-funded MDM initiative, companies will need to address Big Data in an MDM-like fashion. In other words, IT should develop a Darwin-like approach to Big Data governance IT, systematically cataloging this evolving species. This is not a green-light to force the heavy-hand of governance that could throttle innovative uses of Big Data. Organizations need to adapt its centralized governance model to the self-governing world of Big Data. The key to success is to catalog the Big Data sources in a selective

manner, then formulate techniques that help qualify and classify Big Data for business use. Much like IT's initial MDM initiatives, it will take a few iterations for governance to establish the proper blend of policies and processes to support Big Data initiatives. Engaging the business in a rapid but methodical manner should offset initial missteps by Big Data governance, because it demonstrates a strong commitment to key business initiatives.

Beyond symbolic gestures, data governance must take an active role in defining new attributes or metrics that surface from Big Data sources. Within the realm of social media, new customer segmentation concepts, such as influencer score, require examination from a data governance perspective. For instance, is this an attribute of a customer, or is it a new metric measured over time – and if so, which interval should define the snapshot of the score? Moreover, this type of information requires strong data governance, particularly with regards to social media or emails, where privacy is a major concern, and regulations could emerge any day that dictate the standards and use of public Big Data. Governance can also help the business push the boundaries of social media monitoring so it is not exclusive to customer data, but also includes other master data, such as vendor, product, and competitor.

With regards to vendor capabilities, engaging the SaaS model towards Big Data adoption may be appropriate for organizations in the short run. This model has served domain-specific analytic applications well in the past and is already proving viable for Big Data Analytics. The SaaS model is a huge presence in the world of spend analysis, where vendors have pioneered a means of saving companies millions of dollars by collecting various sources of spend data (e.g. Procurement, GL, and AP), enriching the data with third-party sources such as Dun & Bradstreet, and provision spend analysis reports that show spend by product, vendor, or SIC classification, all in an off-premise SaaS platform.

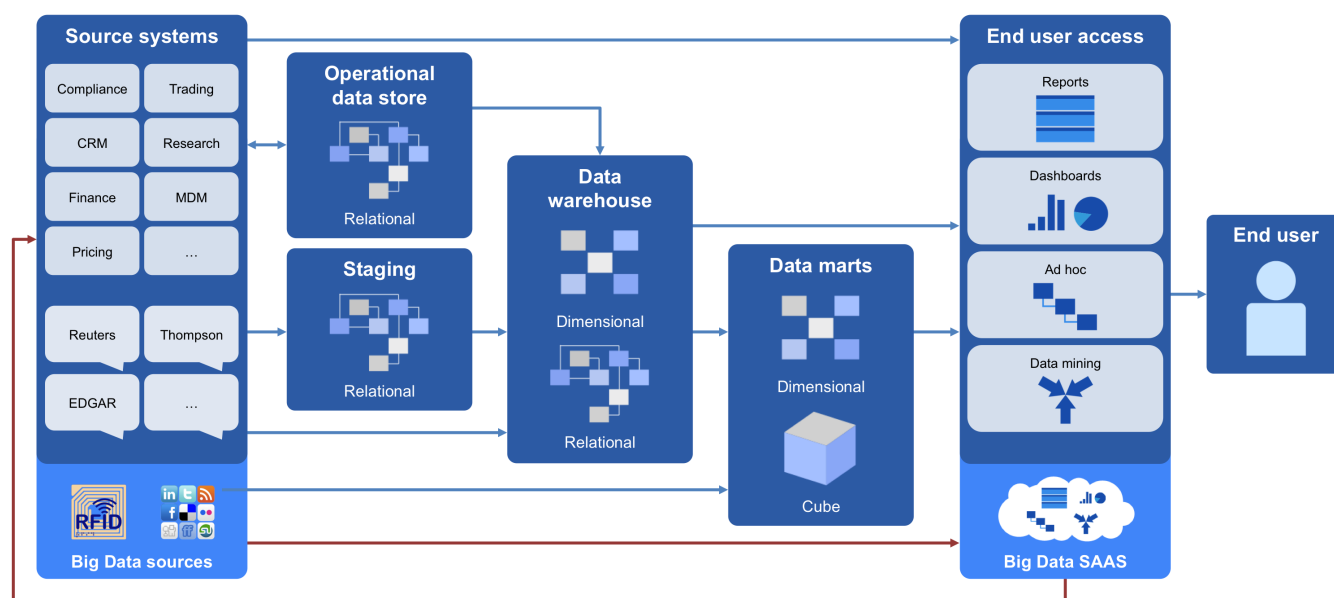
SaaS vendors like Clarabridge, Quantivo, RightNow (Oracle), PostRank (Google), and Radian6 (Salesforce) are doing for social media monitoring what others have done for Spend Analysis. Social media monitoring is a subject highly dependent on Big Interactions – blogs, social media sites, etc. Thus, if social media monitoring is a key business initiative, the SaaS model can help organizations breathe life into its Big Data initiatives. Interestingly, Salesforce acquired Radian6 for \$327 million, an indication of the importance of social media to core CRM and ERP platforms. Oracle followed suit by purchasing RightNow for \$1.5 billion.

Rather than building out an on-premise Big Data environment, organizations can and should leverage the capabilities of innovative SaaS vendors that can help rationalize Big Data sources. One word of advice: before committing to a SaaS vendor, ensure you can use the Big Data from SaaS vendors for on-premise applications, the vendor supports on-premise deployments, or the application can be accessed by onsite apps through standards interfaces (e.g. web services), thereby allowing the reuse and repurposing of Big Data.

Engaging the SaaS model towards Big Data adoption may be appropriate for organizations in the short run. This model has served domain-specific analytic applications well in the past and is already proving viable for Big Data Analytics.

The following illustrates how the SaaS model could compliment an existing warehouse in the short term. In the long term, Big Data will continue to complement, not replace, the existing Business Intelligence infrastructure.

Figure 5: Conceptual Big Data warehouse architecture



The SaaS model makes short-term sense, because companies can deliver value to the business, establish governance policies and procedures that demonstrate business value, and avoid heavy investments in nascent Big Data platforms. In the long-term, this strategy allows the organization internal processes to mature as it pertains to Big Data, and it allows the current generation of Big Data platforms to evolve and the skills in the market to catch-up. It also allows for proofs of concept of Big Data platforms and capabilities in a SaaS model, where prospective buyers can rent first.

For instance, ERP vendors will look to integrate Big Data Sources into their ERP. The acquisition of RightNow by Oracle is an indication of this longer-term shift towards social media monitoring capabilities. Thus if you are an Oracle EBS shop, you may wait for an Oracle social media monitoring OBI App. The timing of this offering will be consistent with the ERP vendors' response to spend analysis, where SaaS vendors like Zycus and Emptoris innovate and lead, and traditional on-premise vendors like Oracle and SAP lag, but catch up to the point where they have competitive social media monitoring solutions that extend their existing ERP platforms.

BI vendors will also make the integration of Big Data platforms transparent (e.g. NoSQL) to developers and end-users so early investments in Big Data training will be short-lived in the long-term. Until enabling Big Data technologies are transparently packaged in an appliance-like fashion, SaaS vendors can help bridge the gap and avoid misguided technology and training investments.

The SaaS model might not be a fit for all Big Data applications in the short-term. For instance, social media monitoring is a specific domain that relies on Big Data and lends itself to SaaS vendors. If a manufacturing company is hoping to mine machine-generated data for an enhanced view of process efficiency, then the potential for a SaaS model is diminished, if not entirely eliminated. In such cases new on-premise Big-data platforms could provide the necessary capabilities.¹²

A Big Data platform isn't necessarily the right answer either. Using the example of the manufacturing company, there are capabilities in existing ERP platforms that provide this operational view of machine sensor data for

manufacturing in real-time.¹³ In such cases the solution may not be a new Big Data platform but operational BI performed by the ERP. The right answer also could leverage an ETL platform to integrate the Big Data source into the existing data warehouse. In all cases, IT must provide the right level of governance and guidance to help answer these critical questions, while not deviating from the existing technical footprint, skill set, or budget.

The bottom line for Big Data is...

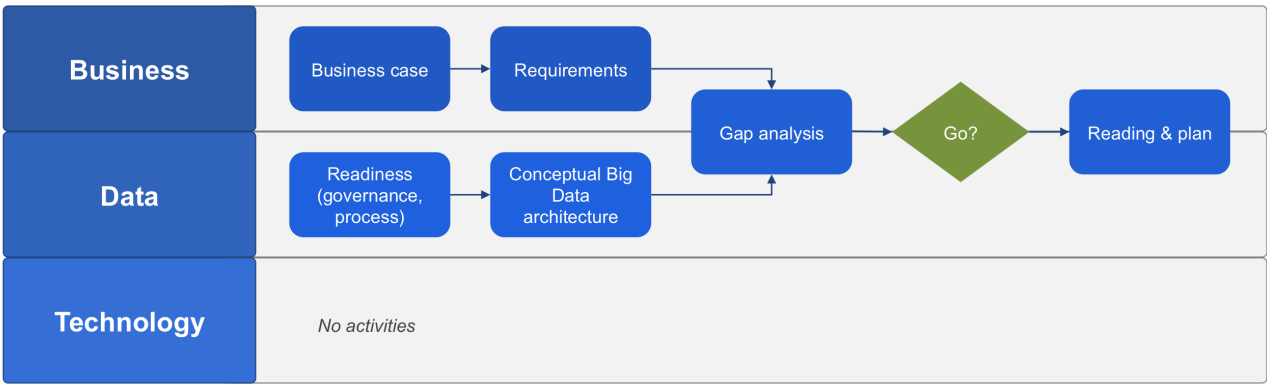
Data has evolved into Big Data. In doing so, the conventional view of the data universe is being challenged. While Big Data can help answer important business questions, organizations must avoid the tendency to invest in the latest technology that is grabbing the headlines. Rather, organizations must adapt internal governance structures and allow Big Data sources to be effectively examined. Companies should also adopt a SaaS-first approach to Big Data technology investments. This two-step approach will enable IT to deliver Big Data business value in the short term, allow its understanding of which Big Data capabilities to invest in, and provide Big Data vendors with the time needed to mature their on-premise platforms.

In all cases, IT must provide the right level of governance and guidance to help answer these critical questions, while not deviating from the existing technical footprint, skill set, or budget.

Collaborative Consulting services help companies make sense of Big Data

Collaborative’s IM services can help formulate the appropriate Big Data strategy – a strategy that aligns a company’s strategic business and technology investments, ensures that data governance adapts to meet the challenges of Big Data, prepares MDM and DI platforms for Big Data, ensures sufficient definition of new BI metrics, and can help with Big Data vendor selection.

Figure 6: Big Data readiness assessment and roadmap



Collaborative’s Business Enablement services can help companies establish a Big Data business case, and assess the OCM impact of Big Data.

References

1. Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simplified data processing on large clusters," Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, December 2004 (http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en/us/archive/mapreduce-osdi04.pdf)
2. http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation
3. <http://www.forbes.com/sites/benkerschberg/2011/11/01/manufacturing-moneyball-using-big-data-and-business-intelligence-to-spur-operational-excellence>
4. IDC, "The big noise around big data", Marc Alvarez, April 3, 2012
5. The DAMA Guide to The Data Management Body of Knowledge, page 4)
6. <http://communicopia.com/insights/four-models-for-managing-digital>
7. <http://radar.oreilly.com/2011/02/big-data-fraud-protection-payment.html>
8. <http://online.wsj.com/article/SB10001424052970204138204576598942105167646.html>
9. USGS Twitter Earthquake Dispatch, <http://twitter.com/USGSted>
10. Descriptions from CrunchBase
11. <http://quantivo.com/blog/top-5-reasons-not-use-hadoop-analytics>
12. Example: Teradata's Manufacturing Analytics Platform
13. <http://www.oracle.com/us/products/applications/057127.pdf>

About Collaborative

Collaborative Consulting is dedicated to helping companies optimize their existing business and technology assets. The company is committed to building long-term relationships and strives to be a trusted partner with every client. Founded in 1999, Collaborative Consulting serves clients from offices across the United States, with headquarters in Burlington, Massachusetts.



© 2012 Collaborative Consulting
877-376-9900
www.collaborative.com