WHITE PAPER

Big Data: What It Is and Why You Should Care

Sponsored by: AMD

Richard L. Villars Carl W. Olofson
Matthew Eastwood
June 2011

EXECUTIVE SUMMARY

Data creation is occurring at a record rate. In 2010, the world generated over 1ZB of data; by 2014, we will generate 7ZB a year. Much of this data explosion is the result of a dramatic increase in devices located at the periphery of the network including embedded sensors, smartphones, and tablet computers. All of this data creates new opportunities to "extract more value" in human genomics, healthcare, oil and gas, search, surveillance, finance, and many other areas. We are entering the age of "Big Data."

IDC believes organizations that are best able to make real-time business decisions using Big Data solutions will thrive, while those that are unable to embrace and make use of this shift will increasingly find themselves at a competitive disadvantage in the market and face potential failure.

Big Data technologies describe a new generation of technologies and architectures, designed so organizations like yours can economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis. This world of Big Data requires a shift in computing architecture so that customers can handle both the data storage requirements and the heavy server processing required to analyze large volumes of data economically.

INFORMATION EVERYWHERE, BUT WHERE'S THE KNOWLEDGE?

Organizations rely on a growing set of applications to communicate with and provide services/products to today's demanding consumer and business communities:

- They are collecting, storing, and analyzing more granular information about more products, people, and transactions than ever before.
- They rely on email, collaboration tools, and mobile devices to communicate and conduct business with customers and business partners.
- They are creating, receiving, and collecting machine and sensor-generated messages, sometimes at very high volumes, and are driving operations and business processes from that message data.

Global Headquarters: 5 Speen Street Framingham, MA 01701 USA P.508.872.8200 F.508.935.4015 www.idc.com

The growth in the number, size, and importance of information assets is not limited to just large government agencies, large enterprises, or Internet Web sites. A wide variety of organizations, ranging from small and medium-sized businesses (SMBs) to large enterprise and government agencies, are dealing with a flood of data as they and their customers:

- Digitize business records and personal content (including the generation of ever-larger numbers of photos, movies, and medical images) driven by continued advancements in device features, resolution, and processor power
- Instrument devices (e.g., set-top boxes, game systems, smartphones, smart meters), buildings, cities, and even entire regions to monitor changes in load, temperatures, locations, traffic patterns, and behaviors
- Address governance, privacy, and regulatory compliance requirements that complicate the retention of business information

Much of this data creation is occurring outside of the datacenter at the edge of the network, the result of a proliferation of embedded sensors and mobile devices. All of this data creates aggregation and analytic opportunities using tools that leverage multicore server architectures and their associated big memory footprints frequently deployed at significant horizontal scale.

This data proliferation also drives demand for centralized storage capacity to maximize the control and usefulness of collected information. To put the pace and newness of this data explosion into context, organizations around the world installed 6.1EB of disk storage capacity in 2007. By 2010, annual new installations were 16.4EB, and by 2014, new capacity installed will reach 79.8EB.

In the past, the main data challenge for most organizations was enabling/recording more and faster transactions. Today, much of the focus is on more and faster delivery of information from scale-out cloud computing clusters (e.g., documents, medical images, movies, gene sequences, data streams, tweets) to systems, PCs, mobile devices, and living rooms. The challenge for the next decade will be finding ways to better analyze, monetize, and capitalize on all this information. It will be the age of Big Data.

IDC believes that organizations that are best able to make real-time business decisions using Big Data streams will thrive, while those that are unable to embrace and make use of this shift will increasingly find themselves at a competitive disadvantage in the market and face potential failure. This will be particularly true in industries experiencing high rates of business change and aggressive consolidation.

So, What Is Big Data?

Big Data is about the growing challenge that organizations face as they deal with large and fast-growing sources of data or information that also present a complex range of analysis and use problems. These can include:

- Having a computing infrastructure that can ingest, validate, and analyze high volumes (size and/or rate) of data

- Assessing mixed data (structured and unstructured) from multiple sources
- Dealing with unpredictable content with no apparent schema or structure
- Enabling real-time or near-real-time collection, analysis, and answers

Big Data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis.

Big Data Sources: What and How Much?

The sources of data growth that are driving Big Data technology investments vary widely. Some represent entirely new data sources, while others are a change in the "resolution" of existing data generated.

New Big Data Sources: Industry Digitization

New data sources for Big Data include industries that just recently began to digitize their content. In virtually all of these cases, data growth rates in the past five years have been near infinite, since in most cases it started from zero. Industries include:

- **Media/entertainment:** The media/entertainment industry moved to digital recording, production, and delivery in the past five years and is now collecting large amounts of rich content and user viewing behaviors.
- **Healthcare:** The healthcare industry is quickly moving to electronic medical records and images, which it wants to use for short-term public health monitoring and long-term epidemiological research programs.
- **Life sciences:** Low-cost gene sequencing (<$1,000) can generate tens of terabytes of information that must be analyzed to look for genetic variations and potential treatment effectiveness.
- **Video surveillance:** Video surveillance is still transitioning from CCTV to IPTV cameras and recording systems that organizations want to analyze for behavioral patterns (security and service enhancement).
- **Transportation, logistics, retail, utilities, and telecommunications:** Sensor data is being generated at an accelerating rate from fleet GPS transceivers, RFID tag readers, smart meters, and cell phones (call data records [CDRs]); that data is used to optimize operations and drive operational business intelligence (BI) to realize immediate business opportunities.

Consumers are increasingly active participants in a self-service marketplace that not only records the use of affinity cards but can increasingly be combined with social networks and location-based metadata, which creates a gold mine of actionable consumer data for retailers, distributors, and manufacturers of consumer packaged goods.

A less obvious industry, but equally interesting, is the legal profession. "Discovery" of electronic records is quickly usurping "discovery" of paper records from both the collection and the review perspective. Leading ediscovery companies are handing terabytes or even petabytes of information that need to be retained and reanalyzed for the full course of the legal proceeding.

Social media solutions such as Facebook, Foursquare, and Twitter are the newest new data sources. A number of new businesses are now building Big Data environments, based on scale-out clusters using power-efficient multicore processors like the AMD Opteron 4000 and 6000 Series platforms, that leverage consumers' (conscious or unconscious) nearly continuous streams of data about themselves (e.g., likes, locations, opinions). Thanks to the "network effect" of successful sites, the total data generated can expand at an exponential rate. One company IDC spoke with collected and analyzed over 4 billion data points (Web site cut-and-paste operations) in its first year of operation and is approaching 20 billion data points less than a year later.

Data growth patterns like these exceed the capabilities of Moore's law and drive the need for hyperscale computing infrastructures that balance performance with density, energy efficiency, and cost.

By definition, the rate of Big Data growth exceeds the capabilities of traditional IT infrastructures and represents largely "greenfield" computing and data management problems for customers. Modular deployment of servers at hyperscale is often the preferred approach to best meet the economic challenges associated with Big Data. As an example, Facebook recently announced its Open Compute Project for both datacenter design and server design. By sharing its modular server design with the market, Facebook hopes to drive down the cost and complexity of hyperscale server deployments aimed at workloads such as Big Data analytics.

Expanding Big Data Sources: Looking Deeper and Further

The other main source of data growth driving Big Data technology investments results from a change in "resolution":

- **Financial transactions:** With the consolidation of global trading environments and the greater use of programmed trading, the volume of transactions that need to be collected and analyzed can double or triple in size, while the transaction volumes can also fluctuate much faster, much wider, and much more unpredictably, and competition among firms forces trading decisions to be made at ever smaller intervals.

- **Smart instrumentation:** The use of intelligent meters in "smart grid" energy systems that shift from a monthly meter read to an "every 15 minute" meter read can translate into a multi-thousandfold increase in data generated.

One of the most interesting use cases is the current evolution of mobile telephony. Until quite recently, the main data generated was essentially limited to caller, receiver, and length of call. With smartphones and tablets, additional data to harvest includes geographic location, text messages, browsing history, and (thanks to the addition of accelerometers) even motions.

Big Data Adoption: Who, Where, and When?

Assessing the range of industries and use cases where Big Data has a role to play is still very much a work in progress. When it comes to the pace of adoption of Big Data practices, culture will play a critical role in many industries. Clearly, organizations that have been generating or dealing with large data sources for decades will be early movers. This group includes many government organizations (e.g., intelligence as well as earth/space/life sciences).

While the specific catalyst for Big Data deployments may vary from customer to customer, it is likely that each deployment has a common trigger point. That is, a tipping point has been reached — in terms of data growth, rate of change, and/or the need for new insights. One services company that recently deployed Big Data recalls:

> I would say that our data volume wasn't that massive, but the rate of change was killing us, where the data volumes were practically doubling every month. Trying to keep up with that growth was an extreme challenge to say the least. Systems weren't designed to deal with that kind of growth, and it was keeping my operations people up all night long trying to deal with all the things that were breaking and straining under the load.
>
> [What attracted us to Hadoop] was its predictability and scalability. For example, budgeting was always a challenge for [us]. How much new hardware am I going to need? How much is it going to cost? With Hadoop, I was able to pretty well plot it out on a linear graph that said when our events monitor reaches this level, this is exactly how much it's going to cost me for hardware ... — Cameron Befus of Tynt.

Other early adopters will come from the classic commercial high-performance computing (HPC) spaces (e.g., oil exploration, design/simulation). AMD led the introduction of 64-bit x86 processor technology with the AMD Opteron processor in April 2003 and subsequently led the introduction of multicore processing to the x86 ecosystem. As a result, many organizations have embraced these computing capabilities as they seek to unlock the power of pervasive data and have developed increasingly hyperscale computing infrastructures to capture, analyze, and deliver actionable business activities using the latest generation of x86 server technologies. This is particularly true for users looking to make use of greenfield Big Data streams.

Another industry that is already a fast technology adopter is financial services. Aside from the analysis of trading patterns for portfolio management, organizations in this sector are also looking to Big Data solutions run on large scale-out clusters composed of multicore x86 processor technology to meet international money laundering and fraud prevention standards such as Basel II.

Many would assume that the telecommunications and retail industries, which certainly care about better/faster analysis of data, would be early adopters, but this is where organizational culture may begin to lead to industry stratification. Both of these industries already make major investments in traditional data warehouse and

business intelligence solutions that often leverage traditional SMP "scale up" server architectures. For them, Big Data will be about both exponential expansion in data sources and faster use of the results (e.g., integration into the logistics system).

Some larger players will seek to graft Big Data environments into existing domains leading to some hybrid infrastructures that leverage both legacy data warehousing and business intelligence coupled with the next general Big Data analytics using scale-out computing infrastructures. Concurrently, IDC expects to see a flowering of smaller competitors or outside "consultants" launching greenfield environments to leapfrog the competition as time-to-market considerations take hold.

Some of the most interesting, but also most challenged, industries when it comes to Big Data adoption will be utilities and content service providers (e.g., cable TV, mobile carriers). These communities (with assists from related companies such as video gaming system and appliance manufacturers) are building out Big Data generating fabrics. Their opportunity now is to figure out how to handle and then do something with all that data, despite the fact that from a cultural standpoint data guardianship and use were much less in the past.

An additional hurdle for these industries is that it isn't enough to just get the "answers" from Big Data platforms. They also need to implement automated response systems (e.g., automated power management or "in game" ad placement) that will ultimately be the foundation of their business models.

Big Data Value: What's in It for Me?

Regardless of industry or sector, the ultimate value of Big Data implementations will be judged based on one or more of three criteria:

- **Does it provide more useful information?** For example, a major retailer might implement a digital video system throughout its stores, not only to monitor theft but to implement a Big Data system to analyze the flow of shoppers — including demographical information such as gender and age — through the store at different times of the day, week, and year. It could also compare flows in different regions with core customer demographics. This move makes it easier for the retailer to tune layouts and promotion spaces on a store-by-store basis.

- **Does it improve the fidelity of the information?** For example, IDC spoke to several earth science and medical epidemiological research teams using Big Data systems to monitor and assess the quality of data being collected from remote sensor systems; they are using Big Data not just to look for patterns but to identify and eliminate false data caused by malfunctions, user error, or temporary environmental anomalies.

- **Does it improve the timeliness of the response?** For example, several private and government healthcare agencies around the world are deploying Big Data systems to reduce the time to detect insurance fraud from months (after checks have been mailed and cashed) to days (eliminating the legal and financial costs associated with fund recovery).

Could Big Data Improve Usefulness, Fidelity, and Timeliness?

A still theoretical but timely example of all three criteria — usefulness, fidelity, and timeliness — would be the use of Big Data technologies in conjunction with accelerometer data feeds from smartphones. Such a system would be able to calculate in near real time the impact/intensity of an earthquake at the bearer's location. Today, geologists have to make inferences based on the intensity at the point of origin, which may be miles underground. These inferences can take hours to solidify. This improvement would make disaster response much more effective in the critical early stages.

BIG DATA DEPLOYMENT: A PRACTICAL GUIDE

If you did a little investigating inside some of your own business units, you would likely be surprised what is already there in terms of Big Data projects. Thanks to a sustained and dramatic decline in the costs of compute power, as exemplified with the AMD Opteron 4000 and 6000 Series processors, memory, and storage capacity (along with new data handling techniques like Hadoop and memcached), it is possible for some bright staff in your organization to effectively deal with many data variety, volume, and velocity problems. Even better, they can do so while still being able to take advantage of the knowledge (get value).

The key questions for you as CIO, IT architect, or business executive are:

- How big of an impact can Big Data have on my IT environment?
- How do we get started and then what will it take to set up an enterprise-class Big Data environment?
- What technological and organizational challenges might inhibit our ability to get the maximum value from Big Data investments?

How Do You Get Started?

Barriers to Big Data adoption are generally cultural rather than technological. In particular, many organizations fail to implement Big Data programs because they are unable to appreciate how data analytics can improve their core business. One the most common triggers for Big Data development is a data explosion that makes existing datasets very large and increasingly difficult to manage via conventional database management tools. As these data sets grow in size — typically ranging from several terabytes to multiple petabytes — businesses face the challenge of capturing, managing, and analyzing the data in an acceptable time frame.

According to Tynt's Cameron Befus:

> A couple thoughts on getting started: You have to take advantage of the training out there because it's a new paradigm and you need to adjust your thinking. Secondly, DevOps, which is an integration of our development and our operations teams, is critical. These two teams work side by side for projects like this; it's absolutely necessary, one group cannot be successful without the other.

Business executives need to improve their ability to convey complicated insights to the organization and drive an effective action plan from the data analysis process. When getting started, it is helpful to think of the following:

- Identify a problem that business leaders can understand and relate to, one that commands their attention.
- Don't get too focused on the technical data management challenge. Be sure to allocate resources to understand the uses for the data inside the business.
- Define the questions needed to meet the business objective and only then focus on discovering the necessary data.
- Understand the tools available to merge the data and the business process so that the result of the data analysis is more actionable.
- Build a scalable x86 infrastructure that can handle growth of the data. You can't do good analysis if you don't have enough computing power to pull in and analyze data. Many folks get discouraged because when they start the analysis process, it is slow and laborious.
- Identify technologies that you can trust. A dizzying variety of open source Big Data software technologies are out there, and most are likely to disappear within a few years. Find one that has professional vendor support, or be prepared to take on permanent maintenance of the technology as well as the solution (probably not a good idea) for the long run. Hadoop is one that seems to be attracting a lot of mainstream vendor support.

  That said, choose a technology that fits the problem. Hadoop is best for large, but relatively simple, dataset filtering, converting, sorting, and analysis. It is also good for sifting through large volumes of text. It is not really an environment for ongoing persistent data management, especially if structural consistency and transactional integrity are required.
- Be aware of changing data formats and changing data needs. For instance, a common problem faced by organizations seeking to use BI solutions to manage marketing campaigns is that those campaigns can be very specifically focused, requiring analysis of data structures that may be in play for only a month or two. Using conventional relational database management system (DBMS) techniques, it can take several weeks for database administrators to get a data warehouse ready to accept the changed data, by which time the campaign is nearly over. A MapReduce solution, such as one built on a Hadoop framework, can reduce those weeks to a day or two. Thus it's not just volume but variety that can drive Big Data adoption.

According to Tynt's Cameron Befus:

> Finding expertise is a challenge but we see signs of improvement. I went to Hadoop World in October 2009 and there were a little over 100 attendees. We were there again last year (October 2010) and this time there were just a little over 1,000 attendees. Everybody

> was still saying that they were hiring anybody that knows anything about this, but a 10-fold growth in one year is fantastic and we see it everywhere. We're talking to people all the time that are saying, "Hey, we want to give this a try."

As organizations work to extract competitive business value — and ultimately revenue — from a growing sea of data, Big Data implementations leverage diverse sets of distributed semi-unstructured and unstructured data types, which frequently start with mathematics, statistics and data aggregation efforts. Big Data analytic software is increasingly deployed on massively parallel clusters leveraging an open source (Apache Hadoop) framework, distributed file systems, distributed databases, MapReduce algorithms, and cloud infrastructure platforms.

Data is expanding faster than the capabilities of Moore's law, and this requires new thinking regarding computing architectures. A world of Big Data requires a shift in computing architecture so that customers can handle both the data storage requirements and heavy server processing required to analyze large volumes of data economically.

Hyperscale server architectures generally consist of thousands of "nodes" containing several processors and disks connected by a high-speed network. Such configurations are rapidly becoming the standard for data-intensive analytic workloads. These systems provide the necessary storage capacity and the raw computing power required to analyze the data and respond to data queries from remote users. Successfully leveraging a hyperscale cluster requires new software algorithms designed to deliver scalability, reliability, and programmability despite the commodity componentry generally constituting the core compute nodes. Finally, the unpredictability of Big Data workloads that are constantly scaling up and down based on data creation patterns will lead users to consider public cloud infrastructures for some of this analysis.

Big Data and the Enterprise: Getting from Science Project to Enterprise-Class Solution

As an organization makes the transition from Big Data as a "junior science project" to Big Data as a core business resource, concerns about the impact on current and future datacenters will increase. Today, and for quite some time, the IT architectural approach used in clustered environments such as a large Hadoop grid is radically different from the converged and virtualized IT environments driving most organizations' datacenter transformation strategies. They have different server/storage configurations and different environmental (power and HVAC) profiles.

Customers will increasingly deploy modular computing infrastructures optimized at the chassis and rack level for optimum processing, memory, I/O, and storage performance. Capacity will be delivered in partial and full rack increments and via hosted offering in the cloud.

In larger enterprises (especially those with global reach), IDC expects to see the emergence of separate datacenters for the workloads. Transaction and content delivery environments, used in traditional Web serving and online transaction

processing, are sensitive to latency, response time, and availability variations, so some geographic dispersion in multiple datacenters makes sense. Conversely, for most Big Data environments, concentration of data streams and compute resources makes more sense for both performance and telecommunications cost reasons. Moving the data in bulk is prohibitive, but moving the much smaller "answers" is not.

In this scenario, individual datacenters become more homogenous, but there will also be an increase in specialty datacenters as some customers elect to deploy clusters at a scale that leverages optimized rack systems and blades with increasingly dense compute and large memory configurations. These differ from the general-purpose servers that are more familiar in today's IT environment, which has been optimized to run a wide range of often disparate workload. In this case, the sheer magnitude of the Big Data problem lends itself to a purpose-built computing infrastructure optimized for Big Data analytic requirements.

From a software perspective, Hadoop is the most commonly used of the new open source Big Data software technologies, but it addresses a limited set of use cases. Other technologies associated with capabilities such as graph databases, recoverable data sharing, and heterogeneous (mixing data formats, and mixing structured and unstructured or semi-structured data) transaction processing are less mature, enjoy less commercial vendor support, and are more volatile.

In short, Hadoop has already begun the move from "science project" to production solutions, whereas the other technologies are still evolving. Do you need a solution right now, and can't wait for the technology to mature? Are you ready to accept the risk of highly changeable technology, and assume responsibility for its maintenance? If so, you could examine Dynamo, Voldemort, Cassandra, and a host of other emerging Big Data software technologies. Otherwise, if your needs can be met by Hadoop (or a similar MapReduce platform) today, and you prefer the security of a supported technology, there are vendors out there that may be able to meet your needs.

Big Data and Cloud: Beginning of a Beautiful Friendship?

The concentration requirement will also make the cloud a critical part of the Big Data picture.

Conceptually, the cloud is about dispersion of computing and storage resources, but in reality, the organizations building the underlying cloud infrastructure are concentrating IT and telecommunications resources to more efficiently deliver applications and information. They are also playing a role in every facet of the Big Data space:

- They will be among the most important collectors/forwarders of data streams and content.
- They will be among the most aggressive users of Big Data systems to run their own businesses.

- They will be in the position to enable Big Data (through simple, temporary provisioning of large compute and data pools) use by technically savvy, but resource constrained, organizations.

Cloud-based Big Data platforms will make it practical for smaller engineering and architectural firms to access massive compute resources for short, semi-predictable time periods without having to build their own Big Data farms.

So, How Big Is Big Data?

Data creation is occurring at a record rate. In fact, IDC's Digital Universe Study predicts that between 2009 and 2020 digital data will grow 44-fold to 35ZB per year. It is also important to recognize that much of this data explosion is the result of an explosion in devices located at the periphery of the network, including embedded sensors, smartphones, and tablet computers. All of this data creates new opportunities for data analytics in human genomics, healthcare, oil and gas, search, surveillance, finance, and many other areas.

Given the scale and scope of developments, one question IDC is frequently asked is, "How big is Big Data today, and how big will it be in the next five years?"

Answering these questions remains a complex task. As already noted, Big Data is not about a single technology, a single architecture, or a single use case. IDC's research on the high-performance computing market (the birthplace of many Big Data ideas) as well as IDC's tracking of server and storage hardware sales to large content depots and cloud service providers (some of today's most aggressive Big Data innovators) can provide some early indicators:

- Spending on server and storage hardware for HPC workloads was $12.4 billion in 2010, and this is projected to grow to $17.2 billion by 2014 (if just 30% of this spend was attributed to Big Data projects with a "real time" emphasis, that would translate into $3.7 billion in spending in 2010). Public cloud and content depots (e.g., Facebook, YouTube, Flickr, iTunes) accounted for 28.8% of all enterprise storage capacity shipped in 2010 and will account for 48.6% of capacity shipped in 2014 (if just 5% of that capacity is used for Big Data processing as opposed to content archiving and delivery, that would still be over 250PB of storage capacity deployed in 2010).

- In 2011, IDC estimates that $14.7 billion will be spent on server and storage hardware for decision support workloads including data warehousing and data analysis. This is more that 17% of all server and storage hardware spending for all workloads deployed across the market.

- Server suppliers shipped more than 51 million processor cores in 2010, which is a 300% increase in capability in just the past five years, while during the same time period, server units increased only 8%.

IDC estimates that scale-out homogenous workloads favoring hyperscale designs optimized for the density and scale requirements of Big Data, Web infrastructure, and HPC workloads will drive 1.9 million server shipments in 2015. This is more than 21%

of all server units shipped, up from 13% in 2010.These data points focus on some of the biggest and most sophisticated environments, but IDC believes that a growing number of organizations in sectors such as retail marketing, securities, media, design, and life sciences are also deploying Big Data environments. Antecedents that focus on the potential impact on specific organizations include:

- A multinational bank indicated a Big Data environment for meeting international money laundering mandates may account for over 20% of server cores deployed in its datacenters over the next four years.
- A social media and Web site monitoring company (<50 employees) is currently running a 200-node Big Data grid that accounts for about 95% of its server and storage assets.
- An electrical utility company is running a couple of racks of blade servers in a grid configuration deploying a Hadoop application that receives smart meter readings in bulk and identifies potential anomalies so that service personnel can fix transformers before they fail, instead of waiting for customers to call and complain that they have no power.

Challenges to Greater Use of Big Data Solutions

Integration with Existing Analytic Environments and Pace of New Application Development

Organizations are struggling to understand the opportunity information provides through advanced analytics. Organizations with high rates of change in their business think about business operations differently. They depend upon data analytics for a wide range of business decisions, often using data analysis to develop business strategies. In short, these users make data-based decisions both more efficiently and at a faster speed than peers typical of their industry.

Traditional business intelligence systems have historically been centrally managed in an enterprise datacenter with the scalable server and high-performance storage infrastructure built around a relational database.

Already, leading vendors that deliver or support BI solutions are actively working to integrate Hadoop functionality into their offerings. Informatica, IBM, Oracle, SAS, and SAP BusinessObjects have all announced efforts along these lines. For the most part, this involves using MapReduce functionality to accept large volumes of data from various sources (in various formats), distill the BI value, generate some reporting, and produce a formatted output dataset that can be loaded into a data warehouse for further processing. In this way, the Hadoop environment is treated as a complementary technology to that of the data warehouse.

Data/Information Security/Privacy

The relatively less structured and informal nature of many Big Data approaches is their strength, but it also poses a problem: if the data involved is sensitive for reasons of privacy, enterprise security, or regulatory requirement, then using such approaches

may represent a serious security breach. Database management systems support security policies that are quite granular, protecting data at both the coarse and the fine grain level from inappropriate access. Big Data software generally has no such safeguards. Enterprises that include any sensitive data in Big Data operations must ensure that the data itself is secure, and that the same data security policies that apply to the data when it exists in databases or files are also enforced in the Big Data context. Failure to do so can have serious negative consequences.

Internal IT Challenges

Big Data also pose a number of internal IT challenges. Big Data buildouts can disrupt current datacenter transformation plans. The use of Big Data pools in the cloud may help overcome this challenge for many companies.

Big Data deployments also require new IT administration and application developer skill sets, and people with these skills are likely to be in short supply for quite some time. You may be able to retrain some existing team members, but once you do, they will be highly sought after by competitors and Big Data solutions providers.

The biggest challenge is the cultural challenge. Today, many of these Big Data projects are best described as "junior science projects" with a small core of servers and storage assets. They aren't the next iteration of a Google-like compute grid, at least not yet. From a business and an IT governance standpoint, however, these kinds of "junior science projects" can quickly turn into the next "Manhattan project" with companywide and industrywide business, organizational, and legal consequences.

ESSENTIAL GUIDANCE: ROLE OF THE CIO AND THE IT ORGANIZATION

Big Data creates infrastructure challenges that extend from data creation to data collection to data storage and analysis. On the analysis side, Big Data favors computing clusters often consisting of hundreds or even thousands of nodes. Each node typically houses two processors and several disks, and the nodes are connected via a high-speed local-area network. These clusters deliver the computing power and storage capacity necessary to organize the data and execute data analysis and user queries. Unlike HPC systems, which generally focus on maximizing raw computing power, Big Data clusters are generally designed to maximize energy efficiency, compute density, and the reliability necessary to manage large data sets. This is done through software algorithms that provide the scalability, reliability, and programmability necessary to run these applications across large pools of computing resources often deployed at hyperscale, which can be located on premise or in the cloud.

Big Data represents both big opportunities and big challenges for CIOs. Almost every CIO dreams about making IT a more valued asset to the organization. Big Data projects are at the frontier of the business, where most of the most significant business expansion or cost reduction opportunities lie. Taking a lead in Big Data efforts provides the CIO with a chance to be a strategic partner with the business unit.

Because speed is strategically important in most Big Data efforts, it will be tempting for business unit teams to move forward without IT support. Your IT team needs to recognize that it must think differently (as well as quickly) and fight for a seat at the table as Big Data strategies are developed.

CIOs need to both understand what their organization is planning around Big Data and begin developing a Big Data IT infrastructure strategy. In doing so, CIOs need to understand the potential value Big Data represents to their organization and industry. IDC believes that building successful business cases around Big Data can only be accomplished through a tight alignment of critical thinking across both IT and the business. This will require out-of-the-box thinking as well as moving outside the traditional IT comfort zone as traditional data warehousing models may not be appropriate to effectively monetize the Big Data opportunity.

CIOs want to be more involved in the business; Big Data can bring IT to the front and center, so be prepared.

Expect that some smart team will soon be coming to talk with you about this great new idea that will transform your company, your customers, or your community. Welcome to the world of Big Data.

PID# 50277A