IDC
Analyze the Future

WHITE PAPER

The Critical Role of the Network in Big Data Applications

Sponsored by: Cisco Systems

Lucinda Borovick Richard L. Villars

February 2012

EXECUTIVE SUMMARY

Throughout the 1990s, the primary IT challenge for most organizations was enabling and recording more and faster transactions for business productivity. Today, in the age of the Internet, much of the focus is on faster delivery of more information (e.g., documents, medical images, movies, gene sequences, sensor data streams) to systems, PCs, mobile devices, and living rooms. The challenge for organizations in the next decade will be finding ways to better analyze, monetize, and capitalize on all these information channels and integrate them into their business. It will be the age of Big Data.

In today's IT marketplace, Big Data is often used as shorthand for 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. IDC believes that organizations that are best able to make real-time business decisions using Big Data will gain a distinct competitive advantage over those that are unable to embrace it. This will be particularly true in industries that experience high rates of business change and aggressive consolidation.

For IT organizations, the emergence of Big Data is about more than deploying a new application or software technology, such as Hadoop. It represents a significant new IT domain that, over time, will require new system designs, administrative skill sets, and data management/access/use policies. Most organizations, however, jump-start their efforts with smaller deployments built on existing resources. As Big Data efforts grow in scope and importance, the network (both within the datacenter and across the WAN) will play a critical role in enabling quick, sustainable expansion while also ensuring these systems are linked to existing mission-critical transaction and content environments.

THE "BIG BANG" OF THE DIGITAL UNIVERSE

The current Big Data wave is first and foremost an outgrowth of what has been happening in many other sectors of the IT landscape. Data creation is taking place at an unprecedented rate and is currently growing at over 60% per year. IDC's Digital Universe Study predicts that between 2009 and 2020, digital data will grow 44-fold to 35ZB per year. The collection of this data, both structured and unstructured, opens the door to and mandates development of new approaches and applications grounded in emerging analytics techniques and emerging processing technologies within the Big Data paradigm.

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

In the corporate sector, business records are being digitized and archived to align with governance, privacy, and other regulatory requirements. For example, in the area of healthcare, huge volumes of medical data and images are being generated and stored as new federal mandates for the use of online medical records begin to come online. On the consumer side, via social networks, content is being stored and utilized, including larger numbers of photos, images, and videos, which enables a highly connected society. In addition to these sources of information, the world of online instrumentation is also expanding. This includes the use of smart meters, onsite video recorders, and devices that monitor changes in load, temperature, location, traffic pattern, and behaviors in buildings, cities, and regions. One example is the use of intelligent meters in "smart grid" energy systems that work on the basis of a reading "every 15 minutes," translating into a multi-thousandfold increase in data generated. Each human- or machine-based activity now generates information that was once confined to isolated domains and is available for much broader usage. In essence, the data deluge is continuous and ever growing.

The richness and diversity of these multi-hundred terabyte and even multi-petabyte data sets are creating a host of new opportunities for Big Data analytics in human genomics, healthcare, oil and gas exploration, search, surveillance, finance, and many other areas. IDC's view of the evolution of the Big Data market is shown in Figure 1. Other vertical markets that are likely candidates for Big Data include media and entertainment, life sciences, transportation, retail, utilities, and telecommunications.

FIGURE 1

The Evolution of Big Data Analytics

Phase / Impact	The Old World			The New Era
	Pilot	Departmental Analytics	Enterprise Analytics	Big Data Analytics
Staff Skills (IT)	Little or no expertise in analytics — basic knowledge of BI tools	Data warehouse team focused on performance, availability, and security	Advanced data modelers and stewards key part of the IT department	Business Analytics Competency Center (BACC) that includes "data scientists"
Staff Skills (Business/IT)	Functional knowledge for BI tools	Few business analysts — limited usage of advanced analytics	Savvy analytical modelers and statisticians utilized	Complex problem solving integrated into BACC
Technology and Tools	Simple historical BI reporting and dashboards	Data warehouse implemented, broad usage of BI tools, limited analytical data marts	In database mining, usage of high-performance computing and analytical appliance	Widespread adoption of appliances for multiple workloads; architecture and governance for emerging technologies
Financial Impact	No substantial financial impact; no ROI models in place	Certain revenue-generating KPIs in place with ROI clearly understood	Significant revenue impact (measured and monitored on a regular basis)	Business strategy and competitive differentiation based on analytics
Data Governance	Little or none (skunk works)	Initial data warehouse model and architecture	Data definitions and models standardized	Clear master data management strategy
Line of Business (LOB)	Frustrated	Visible	Aligned (including LOB executives)	Cross-departmental (with CEO visibility)
CIO Engagement	Hidden	Limited	Involved	Transformative

Source: IDC, 2012

Extracting Value from the Digital Deluge: Identifying Business Benefits

While it is clear that IT organizations will need to address the complexities of managing their ever-escalating data environments, what will differentiate one organization from another in the future is the value that they extract from their organizational data using analytics. The benefits of Big Data will cut across every industry sector. Organizations will have the ability to not only harness data of all types as part of making proactive business decisions but also obtain more useful information for business insights, improve the fidelity of existing information through the process of validation, and improve time to decision making.

Big Data presents myriad possibilities for improvements in business value opportunity for enterprise markets, including use cases such as business fraud detection, security in financial services, and retail business planning. More specific examples are as follows:

- A number of life sciences organizations are analyzing data from clinical trials with the goal of identifying potential adverse effects over longer periods, thereby yielding information that may be too difficult to tease out in the trials themselves.

- A number of the leading content delivery network providers employ Hadoop grids to access viewing patterns for video, music, and game software streams to make proactive decisions on IT and WAN bandwidth allocations.
- eBay is using a Big Data application to address fraud arising from PayPal usage. Business fraud detection has the potential to save millions of dollars.
- A number of insurance companies are augmenting their own claims data with third-party databases that include social media and industrywide sources for anomaly detection targeted to identify new patterns of fraud.

ENTERPRISE-CLASS IMPLEMENTATION REQUIREMENTS

Adopting Big Data environments can pose challenges for IT organizations that have spent the past decades designing and running transaction environments based on traditional relational database management systems (RDBMS) and their associated storage/compute infrastructure.

In some cases, organizations can extend existing data warehouse and business intelligence (BI) environments by leveraging new technologies as coprocessing systems to offload certain tasks. In others, especially in "greenfield" environments such as social media or smart grid environments, they may opt to build the solution based on Big Data technologies for both short-term and long-term cost reasons. In new arenas such as gene sequencing and video pattern detection, traditional solutions were never an option.

The data modeling and architectures involved require integration of new sources of data channels and life-cycle management. This in turn spurs reevaluation of compute, storage, and network infrastructure. Key capabilities to consider are:

- The ability to assess mixed data (structured and unstructured) from multiple sources
- The handling of unpredictable content with no apparent schema or structure
- A software, storage, and computing infrastructure that can ingest, validate, and analyze high volumes of data

Network architecture and components must be flexible enough to adapt and integrate the multifunctional needs of Big Data at variable scale.

Changing Traffic Patterns in the Datacenter

As organizations invest in solutions to achieve these goals, they will need to initiate a shift in computing architectures to accommodate both the heavy storage requirements and the server processing required to analyze large volumes of data efficiently, rapidly, and economically.

In many environments, Big Data computing architectures call for compute resources to be placed near the individual data sets (e.g., localized direct attached storage and

compute), which is different from the centralized storage (e.g., SAN attached) associated with transactional applications. This "localized" data/compute model) introduces two distinct variables — complex data life-cycle management and matching of nodal capacity in terms of compute and I/O need for variety of workloads.

One of the critical tasks for IT organizations will be to balance the needs of this new environment with the traditional transaction-oriented RDBMS that will coexist with Big Data systems. The enterprise network used to support these diverse applications must be optimized to provide a strong foundation for volume, velocity, and accessibility.

These networks must also address a traffic flow shift from the server-to-client pattern that characterizes the traditional enterprise or Web server farm to a heavier server-to-server traffic flow across the datacenter network fabric. This horizontal flow also includes links between servers and increasingly intelligent storage systems. Traditional BI systems have historically been centrally managed in an enterprise datacenter with a scalable server and high-performance storage infrastructure built around a relational database. However, Big Data poses its own unique set of compute infrastructure requirements, encompassing the essential functions of data creation, collection, storage, and analysis.

The special processing requirements involved require dedicated IT infrastructure — typically distributed server clusters often consisting of hundreds or even thousands of nodes. Modular deployment of servers at hyperscale is often the preferred approach to best meet the economic challenges involved.

Each node typically houses two processors and several disks, with the nodes connected via an enterprise-class Ethernet network. These clusters deliver the computing power and storage capacity necessary to organize the data and execute data analysis and user queries. The networking infrastructure that connects each of these nodes must be scalable and resilient enough to optimize performance, especially when data is shuffled between them during specific phases of an application, while also still allowing for external access to share data with the application tier and traditional database management systems.

NETWORK ATTRIBUTES FOR BIG DATA APPLICATIONS

Given the anticipated ability of Big Data to enable organizations to innovate in an increasingly competitive marketplace, IDC believes IT managers will be well served by taking a proactive approach to planning their network architecture in support of Big Data analytics. These projects will expand in scope and become inextricably linked with business processes. As a result, the network will serve as the central unifying foundation that enables the interconnectivity between existing investments and newer Big Data projects. Business wants to realize the benefits of Big Data analytics without the downside of duplicative capital investments, skyrocketing operational costs, and the introduction of faulty risk management policies, any of which could sidetrack current best practices in datacenter design. A well-thought-out approach to the datacenter network can proactively mitigate these concerns. IT organizations should start the planning process by identifying the optimal network attributes necessary to support Hadoop clusters or other configurations.

Predictable and Efficient

Big Data analytic software is increasingly being deployed on massively parallel clusters leveraging the Apache Hadoop framework, including MapReduce and the Hadoop Distributed File System (HDFS). Hadoop is now a widely used platform and provides a highly scalable runtime environment for large clusters.

The massive amounts of diverse data being transferred with Big Data applications increase the amount of real-time and workload-intensive transactions significantly. Because hyperscale server architectures may consist of thousands of nodes containing several processors and disks, the supporting network connecting them must be robust enough to ensure this data can move quickly and efficiently.

Traffic patterns are likely to be bursty and variable due in part to uncertainties regarding the amount of data moving over the network at any given time. When a data load task is requested, data must be transferred and distributed to the cluster via the network. Delays in data transfer can be significant unless the requisite network resources are in play. While predictable and consistent latency is important, these applications are not typically nanosecond latency sensitive.

Appropriate line rate performance furthers network efficiency. Rightsizing switch capacity is one of the essential tasks needed to achieve network efficiency. In the current environment, typical network configurations for Big Data are likely to require 1GbE access layer switch capacity. Over the next 12–18 months, as the cost/performance ratio becomes more efficient, 10GbE server connectivity will become more common, and it is possible that some organizations may need to upgrade aggregation switch capacity to 40GbE or even 100GbE.

Holistic Network

Optimized network performance needs to take place within the Big Data domain itself as well as in the connection with the more traditional enterprise infrastructure. The network is an essential foundation for transactions between massively parallel servers within Hadoop or other architectures and between the server cluster and existing enterprise storage systems, whether a DAS, SAN, or NAS arrangement is involved.

The benefits to a holistic network approach include:

- Ability to minimize duplicative costs whereby one network can support all workloads
- Multitenancy to consolidate and centralize Big Data projects
- Ease of network provisioning where sophisticated intelligence is used to manage workloads based on business priorities
- Ability to leverage network staffing expertise across the entire datacenter

Network Partitioning

Another key area affecting network design and implementation has to do with governance or regulatory considerations. For example, a research institution may have multiple projects under way that need to be separated, or a financial company may want to separate Big Data projects from regular transaction-oriented line-of-business IT resources on the network. Logical partitioning enables this separation without adding cost and complexity. In addition, various tasks might also need to be isolated using hard partitioning on the Ethernet switch. This means that tasks are completely separated at the data plane level. For example, data plane separation would be needed to comply with regulations and privacy requirements associated with healthcare applications that may contain sensitive data.

Scale Out

Big Data projects may start small as "junior science projects," but the ability to "scale out" will ensure a seamless transition as projects increase in size and number. In addition, and equally important, network performance and ease of management remain constant as the cluster scales. Because of the demands of machine-to-machine traffic flows, oversubscription should be minimized within the Big Data cluster network.

Unified Ethernet Fabrics

Unified Ethernet fabrics are emerging as the necessary network better suited to the needs of cloud computing and Big Data. IDC defines a unified Ethernet fabric as a flatter and converged network. The converged network reduces the complexity and expense associated with multiple fabrics, separate adapters, and cabling. Additionally, network architects favor flatter designs to maximize network efficiency, reduce congestion, and address Spanning Tree limitations by creating active/active Layer 2 network paths for load balancing and redundancy.

IDC believes that the use of unified Ethernet fabrics is important because, compared with traditional Ethernet fabrics, they can optimize application performance and availability and at the same time reduce costs and complexity. Unified fabrics support Big Data as an additional type of application load. It is an application workload that fits into the unified Ethernet fabric designs rather than demanding an entirely new network design.

Unified Ethernet fabrics enable full link utilization by leveraging multiple paths through the network and continuously determining the most efficient route. Ethernet fabrics offer excellent scalability because virtual chassis architectures provide access to multiple switches while logically managing them as a single device. This creates a pool of virtual switching resources and eliminates the need for manual configuration. This design also provides predictable any-to-any latency and bandwidth for traffic between servers within the Big Data cluster. In addition, the fabric brings a distributed approach to networking that is much more resilient to failures.

REALIZING THE BENEFITS OF BIG DATA WITH CISCO UNIFIED FABRIC

To address the opportunity in Big Data, Cisco is delivering Unified Fabric. Cisco Unified Fabric is a key building block for both traditional and virtualized datacenters and unifies storage and data networking with the intent of delivering seamless scalability, convergence, and network intelligence. As a foundational pillar for Cisco's Unified Data Center Framework, it complements Unified Network Management and Unified Computing with the goal of delivering compelling business value to customers.

Scalability

Unified Fabric's scalability features enable resilient, flexible, cost-effective network growth to match fast-growing, widely distributed Big Data architectures while maintaining high performance and ease of management and design. Unified Fabric also provides consistent any-to-any low-latency, high-bandwidth, and resilient performance as Big Data environments scale from tens to hundreds to thousands of systems.

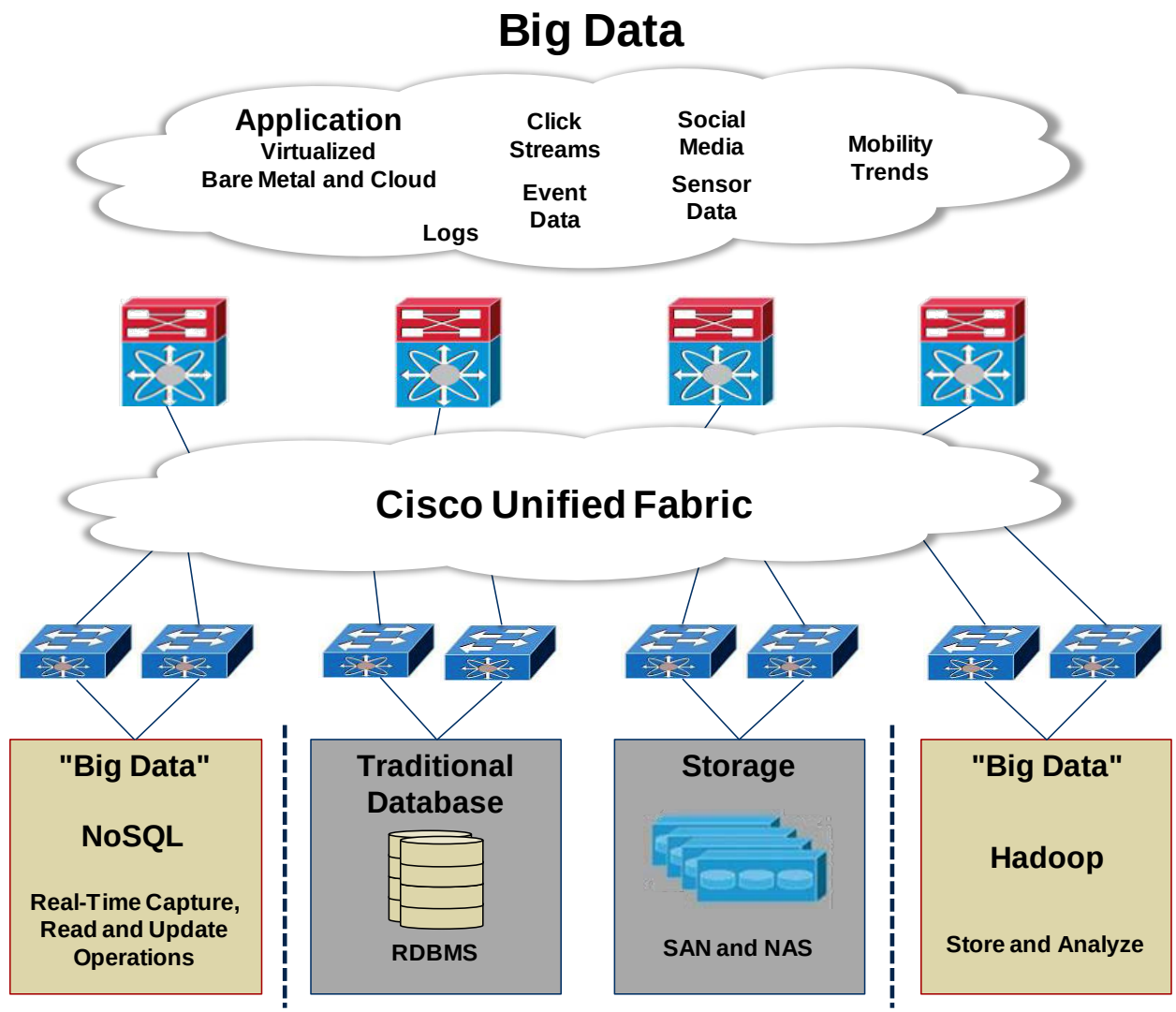
Unified Fabric is intended to support scalability on multiple fronts by dynamically accommodating changing traffic patterns including larger, more complex workloads such as Big Data. It does this by including all network locations in a single extended environment whereby capability scales automatically with changes in the size of the network. Cisco's infrastructure portfolio covering Fabric Extenders as well the Nexus 3000 provides a flexible approach suited for scaled deployment with and integration into the enterprise network, providing connectivity to "structured" data sources. It also continues the company's strategy of embedding policy-based, intelligent services in the network with the goal of faster application deployment through the avoidance of manually making physical infrastructure changes. As Big Data applications grow in size, these capabilities should be of significant benefit.

Convergence

The convergence features of Unified Fabric are designed to enable multiprotocol access and provide a flexible and comprehensive datacenter solution for delivering agile and cost-effective network services to servers, storage, and applications. Unified Fabric helps provide a common fabric for both Big Data and traditional SAN/NAS-based architectures and enables robust connectivity between both. Unified Fabric delivers intelligent services directly into the network fabric. It transparently extends the network to encompass all locations into a single environment with consistent services and policies, which help Big Data networks efficiently and securely connect to other parts of the enterprise (see Figure 2).

FIGURE 2

Big Data Building Blocks into the Enterprise



Source: Cisco Systems, 2012

Network Intelligence

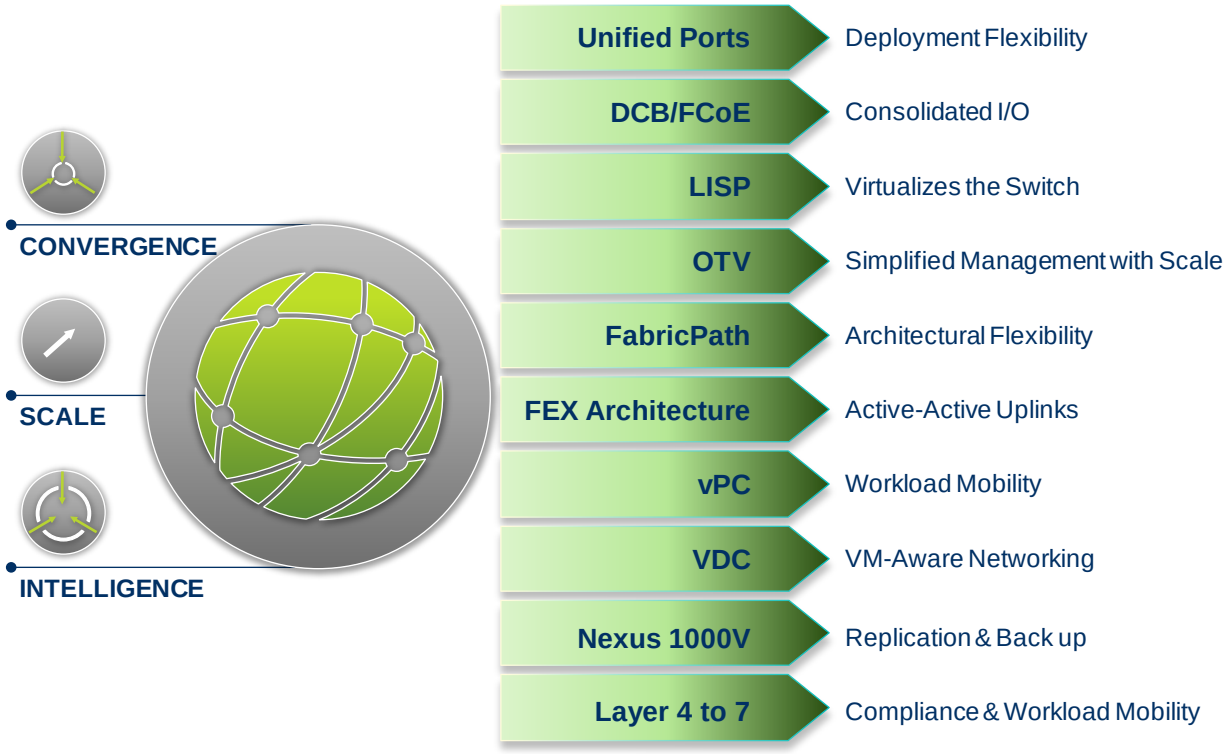
For more traditional application architectures, another benefit of Unified Fabric is the ability to integrate the way server and storage resources are connected, resulting in improved network intelligence. In terms of storage, it has the ability to carry diverse traffic, including Fibre Channel, FCoE, iSCSI, and NAS, over an Ethernet connection. It can also simplify datacenter networks by converging LANs and SANs via the use of Data Center Bridging (DCB) and unified ports as well as FCoE, which is available on the MDS 9500, Nexus 7000, and Nexus 5000 platforms. This flexibility can enhance the heavy east-west any-to-any traffic flows of a Big Data DAS-based cluster as well as traditional SAN/NAS-based database applications, enabling customers to manage a single network fabric regardless of application requirements or storage designs.

Unified Fabric Benefits for Big Data

The intent of Unified Fabric is a network that scales, is based on a converged Ethernet network, and has the intelligence to take a holistic systems-based approach to foundational network capabilities (see Figure 3).

FIGURE 3

Cisco Unified Fabric



Source: Cisco Systems, 2012

One of the major goals of Unified Fabric is to improve network resiliency and availability. Another is to minimize total cost of ownership (TCO) by reducing capex related to cabling, host interfaces, and switching ports and opex related to considerations such as power, cooling, and floor space. Unified Fabric provides low cost per 1GE port at scale and a smooth transition from 1GE to 10GE as Big Data architectures evolve and demand more from the network, without changing the management model or network topology.

Unified Fabric brings the following benefits to Big Data:

- **Scalability.** As Big Data applications grow in size, the fabric can scale incrementally with the projects. Fabric intelligence is essential to this seamless scale.

- **Multitenant architecture.** Given the applicability of Big Data projects across business units, the fabric's ability to provide a multitenant architecture furthers an organization's ability to leverage the fabric across multiple use cases.

- **Machine-to-machine traffic.** The fabric is designed for machine-to-machine traffic flows with resource buffering that is integral to Big Data infrastructure architectures.

KEY CONSIDERATIONS ON THE PATH TO IMPLEMENTATION

Organizations are struggling to understand the opportunity that Big Data provides through its advanced analytics and the scope of its data reach into multiple sources. However, implementation involves complexities that lie beyond the scope of the current capabilities of many IT departments. This means that the move to Big Data can pose a number of challenges organizationally. From a staffing perspective, for example, Big Data deployments will require new IT administrative and application developer skill sets in areas such as machine learning and statistics. However, staff with the requisite training may be hard to find over the short term. In addition, the relatively less structured and informal nature of many Big Data approaches is their strength, yet less structure and informality pose a number of privacy-related concerns. Like all other corporate data, Big Data requires organizations to take into consideration security, privacy, and governance needs.

Given the expected explosion in Big Data projects, it is critical that IT organizations prepare for a scale-out type of network deployment. The ability to both fit into existing network topologies and ensure that the network provides a seamless path to expand is a paramount concern. IDC believes that organizations need to consider Big Data as a critical but additional workload to be accommodated on a unified fabric. This approach will have downstream benefits along the lines of resiliency, logical separation, ease of management, and cost-effective "pay as you grow" economics.

If your organization has struggled with the concept or implementation of a unified network strategy up until now, the potential for including Big Data applications within your environment should at least serve as an opportunity to revisit your current networking strategy.

CONCLUSION

Big Data represents a major evolutionary step related to the broad objectives of business analytics offering new opportunities for the way information is used in proactive organizational decision making. As pilot and implementation plans are considered, IDC believes that the support needed for terabyte-sized structured and unstructured data set analytics can be addressed at the enterprise level using currently available but customized technology, toolsets, and business infrastructure.

Because Big Data is at the frontier of a new wave of business transformation, IT organizations will need to consider a coordinated approach to planning implementations. Most importantly, they need to develop an IT infrastructure strategy

that optimizes the network as a foundational element within existing unified datacenter environments. IDC believes that well-developed plans for networking support of Big Data should address optimizing the network both within a Big Data domain and in the connection to traditional enterprise infrastructure. Accomplishing this will require a number of considerations, including building architectures that handle availability and resiliency burst at scale while providing unified access to all data models. Addressing this variety of network considerations not only will be fundamentally necessary if customers plan to build a solid foundation for Big Data applications within their organization but also will be a requirement for meeting the concerns imposed by data governance and regulatory requirements.