HITACHI
Inspire the Next

DATA DRIVEN GLOBAL VISION CLOUD PLATFORM STRATE
ON POWERFUL RELEVANT PERFORMANCE SOLUTION CLO
VIRTUAL BIG DATA SOLUTION ROI FLEXIBLE DATA DRIVEN V

WHITE PAPER

Build the Future of Big Data Today

By Hitachi Data Systems

November 2013

Hitachi Data Systems

Contents

Build the Future of Big Data Today

Executive Summary

Leaders are anxious to harness the power of big data analytics to deliver on the promise of business insight and competitive advantage. Whether human, machine-generated or both, the volume, velocity and variety of big data make big demands on the underlying infrastructure. Having the right infrastructure underpinnings can make or break your big data project. This white paper examines the infrastructure needs of big data and the connection between your information infrastructure and the success of your big data projects. It also explores the synergy between the infrastructure needs of big data and the trend toward "software defined" IT environments. It ends with a checklist you can use to gauge your infrastructure readiness for big data projects.

The Big Data Opportunity

Big data is not a trend; it's shorthand for business opportunity. Why is data bigger and why is it a business opportunity now more so than in the past? Three important advancements are underway. First, there is more human-generated data being produced through interactions with social media, mobile and messaging applications. Second, keen interest in monetizing this data has driven innovation around data management and analysis techniques and technologies into high gear. Third, these advancements are shining new light on other types of data that were previously untapped. One of these data types is machine-generated data: Think energy sensors, security logs, health care imaging, and the like. Another is so-called enterprise dark data: data that businesses have retained, but have not yet analyzed.

We are bombarded by factoids. Here's an example: By 2015, global data will reach 7.9ZB! Enter the "3 Vs": volume, velocity and variety. In other words, there is a very, very large amount of data. Both humans and machines are generating it at an unprecedented pace, and it comes in many types and from all directions. The business opportunity is in analyzing this big data to learn something you did not already know or to back up your theories with facts. These analyses promise insights for smarter business decisions, competitive advantage, better customer relationships, and loyalty and product improvements. If you are thinking that this sounds good, you can get started with a solid foundation. The foundation for every big data project is the right infrastructure. The following sections outline the infrastructure essentials for big data.

Big Data and Software-Defined IT: A Good Match

The ideal infrastructure for big data shares common characteristics with the shift toward software-defined IT environments. The challenges around big data in many ways are the impetus for software-defined data centers. According to the 2012 IDC Data Center Infrastructure Survey, data center capacity is slated to grow by more than 50% over the next 5 years. But this new capacity must serve an emerging, diverse set of workloads: Big data analytics, content serving, archiving and mobility together are eclipsing traditional transactional workloads. Because of the business opportunity, big data analytics are playing a major role in emerging architectures.

The data mix has changed along with the workloads. Traditionally, data was predominately structured. In other words, could be organized in rows and columns (think relational databases, bank and credit card transactions, purchase orders, and so forth). Global data is now 80% or more unstructured (think photos, X-rays, videos, emails, tweets, and so forth). Unstructured data rely on file systems to translate sequences of bits or bytes stored on disks in fixed lengths or "blocks" into user- or client-understandable formats. A data block by itself has no knowledge of the file to which it belongs. Structured data stored in databases understand (read and write) blocks directly because they contain their own "filing systems." Both data types must be analyzed, often together and concurrently, to meet big data business objectives.

Big data business opportunities are usually time and event dependent. For example, to improve sales, a business may wish to predict and influence customer behavior by making offers as targeted and personalized as possible. Or, a business may need to identify an anomaly, such as a product defect, quickly, before it causes revenue loss. As such, multiple data sources must come together real time or at least within a designated window of time. More than one application might require access to multiple data sources.

Big data infrastructure must deliver the right type of compute resources when data, applications and end users need them, or business opportunities might be missed. Big data requires a top-down approach to information infrastructure where requests from business owners are translated into compute resources tailored to the job at hand. It requires a new approach to infrastructure, one that is more programmable, or, if you will, software-defined.

The core information infrastructure characteristics shared by the needs of big data and software-defined IT are abstraction, extensibility and automation. Look for these characteristics throughout this white paper. These characteristics are not easy to deliver, but technology is advancing to get us there. Many solutions are already firmly established in today's IT environment. Others, newer to IT, are essential to your big data information infrastructure. The remainder of this white paper is devoted to an overview of these big data infrastructure essentials.

Efficiency and Flexibility

Optimization at every level of the big data stack is an imperative. Big data analytics projects are notoriously unpredictable and can involve very large, complex data sets from many sources. At the infrastructure layer, there are opportunities to inject both efficiency and flexibility that serve the entire stack.

Virtualization

Virtualization is the ***abstraction*** of infrastructure and other IT resources. Server virtualization has become universally accepted as a way to make dramatically more efficient and flexible use of physical servers: Physical servers can be carved into many virtual servers, which can be created nearly instantaneously and deleted when no longer required, making the physical resources available once again.

Storage virtualization brings a similar level of efficiency and flexibility to big data infrastructures. It makes it much easier and more cost-effective to access large volumes of structured and unstructured data. The virtualized storage resources are presented as pools of on-demand data to big data analytics applications. Using ***automation***, the utilization rate of well-managed, virtualized tiered storage can approach 100%, making it very cost-effective.

Likewise, network virtualization or software-defined networks take a number of physical network connections and present them as pools of connections, ***abstracting*** underlying network complexity.

Bottom line: Virtualization addresses big data challenges by allowing your infrastructure to move constantly changing big data workloads flexibly to the right compute and storage resources. It also allows resources to expand quickly and efficiently to meet the volume, velocity and variety of big data.

Performance and Scale

If virtualization is the efficiency and flexibility superhero of big data infrastructure, then performance and scale are the dynamic duo of extensibility. A sometimes-misunderstood feature of big data is that it can start small and grow bigger. Or, its "bigness" can originate with the notion that it must be obtained from many sources and analyzed together. In other words, sometimes the "big" in big data is not the sheer volume of data, but the variety and/or the velocity. Therefore, the ability to scale linearly for capacity and performance is important to big data infrastructure.

Distributed Compute Environments

Distributed computing refers to the ability to spread demanding workloads, like big data analytics, across clusters of compute resources. Each cluster or "node" contains CPU and memory and access to storage. The compute resources can be physical or virtualized. The storage can be local to each node or accessed from a virtualized pool. The result is a collection of independent compute resources that appear to the application or user as a single coherent system.

Analytics applications require performance and the ability to add more data, or capacity, as needed. As such, they thrive in low-latency environments that can scale on demand. Distributed compute clusters present this type of environment, which is highly ***extensible*** without sacrificing performance.

Hadoop is an example of an open source, breakthrough big data management technology that takes advantage of distributed compute environments and the ***extensibility*** they offer. Hadoop breaks big data into manageable chunks. It processes these data chunks in parallel on a distributed cluster and makes the data available to users and applications as output or for further processing.

Converged Infrastructures

By now you know that big data analytics are compute and data-intensive applications. As such, infrastructure proximity and quality of integration matter a great deal. IDC maintains data analytics are one of the business drivers of converged systems.

As discussed, virtualization allows the movement of workloads to the physical infrastructure that can handle their scale and performance demands. But first, the right physical infrastructure and integration with virtualization must be in place. Pre-integration with big data analytics applications and platforms is also a consideration.

Converged infrastructures offer servers, storage, networking, software and management as a single orderable solution. Converged infrastructure systems are pre-integrated, can use pooled resources, and are managed with some level of ***automation***. Therefore, they also hold the promise of lower operating expenses. You do not need to start from scratch with your big data infrastructure.

Object Storage

One of the easiest ways to describe object storage is to start with what it is not. It is not a file system that stores and manages data in a hierarchy. It is not block-based storage that stores and manages data in volumes striped across disk drives. Object stores do augment and complement both of these data storage methods. An object store is akin to a database for unstructured data that holds the object references, plus a distributed file store that holds the user data. In object stores, the details of the underlying file systems are ***abstracted*** from applications, end users and system administrators. What this ***abstraction*** means for big data is the answer to the unstructured data problem described earlier. That is, historically, unstructured data has not been easy to manage or analyze.

Object stores offer tremendous scale: the ability to store and manage billions of objects. Moreover, they incorporate both system (application-supplied) and custom (user-supplied) metadata or "data about data." Since it takes an order of magnitude greater compute cycles, effort and time to search full data versus metadata, the savings are a boon for big data. A system that can readily search big data has made a giant step forward in supporting big data analytics.

High Availability

It is imperative to ensure that your big data infrastructure remains available to your big data platforms and applications at all times. Physical device redundancy and caching methods combined with data replication to a remote site are tried and true methods for high availability. Fault-tolerant distributed file systems help assure continued service despite, and in anticipation of, inevitable system component failures. Not all high-availability methods are created equal, so it is important to understand how high availability is supported within your big data infrastructure.

Data Protection

Data protection is on the critical path for big data projects. Data is the resource that fuels big data analysis, so it must be safeguarded. Protection starts at the physical storage layer with RAID-6 (redundant array of independent disks) protection from disk failures. There are other types of so-called erasure coding algorithms as well.

Modern data protection methods typically use a series of point-in-time data snapshots combined with replication to a remote location. A backup management application can orchestrate snapshots and replication and make the data recoverable from both a recovery point objective (RPO) and a recovery time objective (RTO). Based on the criticality of data, you choose an amount of data that is acceptable to lose. You also choose the amount of time you are willing to wait for data to be restored to your systems. Solutions are offered for a range of RPOs and RTOs, including zero data loss and near-zero time to recover.

If your big data infrastructure uses an object store with a redundant replica capability, this approach can reduce the need for traditional backup.

Hitachi Data Systems Has Answers

Hitachi Data Systems information infrastructure for big data offers virtualized infrastructure platforms, content platforms, big data services and integration with analytics platforms. Hitachi Unified Compute Platform (UCP) Select for the SAP HANA Platform is a converged infrastructure solution that combines storage, server, networking and software management. Unified Compute Platform for business analytics is optimized for the SAP HANA platform for high-speed in-memory analytics.

Hitachi offers the most advanced storage virtualization available with Hitachi Virtual Storage Platform. Hitachi Content Platform (HCP) is an object-based content platform ideal for big data management. It makes content far easier to discover, access and analyze. Both UCP and HCP use application protocol interfaces (APIs) to the cloud so big data from social media and other external sources can be accessed readily.

With a robust virtualization ecosystem, highly resilient architectures and highly evolved data management, Hitachi has the right infrastructure for big data platforms and applications, including analytics engines like Hadoop. Hitachi Consulting can help identify big data business opportunities and help evaluate and architect big data solutions.

Visit www.HDS.com for more information.

Appendix: Checklist for Big Data Infrastructure Readiness

✓ DISTRIBUTED COMPUTING

Distributed compute resources with horizontal scalability and massive parallel processing allow multiple concurrent big data analytics. Distributed computing supports data-intensive operations such as map and reduce, which are used in Hadoop clusters.

Big Data Advantage

Computational power, speed and scale to process, manage and manipulate immense amounts of data.

✓ VIRTUALIZED INFRASTRUCTURE

A virtualized infrastructure incorporates server, storage and network virtualization with application programming interfaces (APIs) to leverage functionality so that it is expertly handled by each layer. For example, workload agility and mobility in the server layer, data performance and persistence in the storage layer and connectivity in the network layer. Storage virtualization provisions storage in efficient, consolidated pools with the right mix of performance and capacity for big data on a petabyte scale.

Big Data Advantage

Increased efficiency, flexibility, resource utilization and consistency of data presentation to big data analytics applications.

✓ CONVERGED INFRASTRUCTURE

A converged infrastructure includes pre-integrated server, storage, network and software. These components may be pre-integrated with or tuned for specific big data applications like in-memory analytics and optimized for big data workloads, such as Hadoop-based or similar NoSQL distributed frameworks. Consider converged infrastructures that are pre-integrated with virtualization technologies. Also consider those that are pre-integrated with automated management features and interfaces to external big data sources, such as the Internet or cloud (RESTful API).

Big Data Advantage

Faster time to big data analysis, reduced complexity and lower infrastructure operating costs.

✓ OBJECT STORE

Object stores offer tremendous scalability, with the ability to store and manage billions of objects. Moreover, they incorporate both system (application-supplied) and custom (user-supplied) metadata or “data about data.”

Big Data Advantage

Massive scalability and ready search for unstructured data, a key requirement for big data analytics. (It takes an order of magnitude greater compute cycles, effort and time to search full data versus metadata.)

✓ HIGH AVAILABILITY

Physical device redundancy and caching methods combined with data replication to a remote site are tried and true methods for high availability. Fault-tolerant distributed file systems help assure uninterrupted service.

Big Data Advantage

Continuous data management and analytics operations.

✓ DATA PROTECTION

Protect data at the physical storage layer with RAID-6 (redundant array of independent disks) or other device-appropriate erasure coding algorithms. Further protect data using a series of point-in-time data snapshots combined with replication to a remote location. A backup management application can orchestrate these and make the data recoverable according to recovery point and recovery time objectives.

Big Data Advantage

Data recoverability from operational failures and disaster.

Hitachi Data Systems

Corporate Headquarters
2845 Lafayette Street
Santa Clara, CA 96050-2639 USA
www.HDS.com community.HDS.com

Regional Contact Information
Americas: +1 408 970 1000 or info@hds.com
Europe, Middle East and Africa: +44 (0) 1753 618000 or info.emea@hds.com
Asia Pacific: +852 3189 7900 or hds.marketing.apac@hds.com

WP-472-A DG November 2013