Table of Contents

Business Case .................................................................................................................. 4
Executive Summary .......................................................................................................... 4
Scope .................................................................................................................................. 4
Audience ............................................................................................................................. 4

Emerging Infrastructure Challenges for Big Data Analytics ............................................ 5

Why Vblock® Systems with EMC Isilon for Hadoop? .................................................. 6

Vblock Systems: A Foundation for Unstructured Data Value ......................................... 8

Big Data and Analytics on Vblock Systems..................................................................... 9

Business Benefits of Big Data Analytics on Vblock Systems.......................................... 10

Vblock Systems Overview .............................................................................................. 10
   VBLOCK SYSTEM 700 FAMILY .................................................................................. 11
   VBLOCK SYSTEM 500 FAMILY ................................................................................ 11
   VBLOCK SYSTEM 300 FAMILY ................................................................................ 11
   VBLOCK SYSTEM 200 FAMILY ................................................................................ 11
   VBLOCK SYSTEM 100 FAMILY ................................................................................ 11
   VCE TECHNOLOGY EXTENSIONS ......................................................................... 11

EMC Isilon Scale-Out NAS for In-Place Hadoop Analytics ............................................ 11

Scale-Out Storage to Meet Big Data Demands ................................................................. 12
Separation of Data from Compute .................................................................................. 13
Native Hadoop Integration with Distribution Portability .................................................. 14
Robust and Efficient Data Protection ............................................................................. 14
High Resilience and Availability .................................................................................... 15
Enterprise-Class Security and Compliance ................................................................... 16
Multi-Protocol Support .................................................................................................... 16

Big Data and Analytic Use Case Examples ................................................................... 16
   Market Intelligence Researcher .................................................................................... 16
   A Leading Managed Hosting Provider ....................................................................... 17
   Telecommunications Leader ....................................................................................... 17
   Pharmaceutical Giant .................................................................................................. 17
   Global Financial Services Provider ........................................................................... 17
   Multi-Channel Retail Chain ....................................................................................... 17
   Major Food Chain ....................................................................................................... 18

Vblock Systems with VCE™ technology extension with EMC® Isilon® ................. 18

Conclusion ....................................................................................................................... 19

Next Steps ......................................................................................................................... 19
Business Case

Executive Summary
Big data analytics is at a crossroads. IT is under increasing pressure from the business to deliver tangible value from unstructured data generated across dozens of sources—web commerce, social media, sensors and devices, geospatial systems, internal and external cloud systems, and more. As a result, IT leaders are strategizing how to move Hadoop-based analytics efforts from experimental sandbox prototypes to production environments that integrate with the overall enterprise IT infrastructure.

The choice of infrastructure systems is critical to maximizing the success of your big data and analytic projects, particularly as information volumes, variety, and velocity continue to soar, and the timely analysis of this data has direct business impact. The commodity servers with direct-attached storage (DAS) systems in place for Hadoop prototypes and pilot projects suffer from numerous limitations, from difficult data ingest and inefficient data duplication to availability, privacy, and security concerns that undermine their suitability for enterprise-class big data and analytics.

This white paper examines the infrastructure challenges and considerations that enterprises face in evolving their Hadoop analytics practices towards production environments. The enterprise-ready Hadoop infrastructure discussed is designed to support corporate mandates to improve operational efficiency, combat fraud and risk, support collaborative partner networks, and increase customer retention and acquisition.

It also details how Vblock® Systems from VCE address enterprise big data requirements, with the option to deploy EMC Isilon network-attached storage (NAS) for scale-out capacity and performance to enable unstructured data management and analytics, and better long-term data retention, governance, and security. Formed by Cisco and EMC with investments from VMware and Intel, VCE is the leading innovator of converged infrastructure that combines compute, storage, networking, and virtualization with a management layer.

With integrated, native Hadoop support, Vblock Systems and Isilon combine to deliver a scale-out infrastructure that enables enterprises to achieve the highest levels of reliability, availability, performance, and security in the analytics production environments.

Scope
This paper explores the challenges confronting IT professionals in incorporating Hadoop-based big data analytics initiatives into a broader enterprise infrastructure. It discusses the use of converged infrastructure with scale-out capacity and performance, and examines VCE Vblock Systems and EMC Isilon scale-out NAS and foundational elements of a modern infrastructure for enterprise Hadoop.

Audience
This document is intended for large enterprises and service providers that are evaluating a converged infrastructure with scale-out capacity for big data analytics moving to enterprise Hadoop.
Emerging Infrastructure Challenges for Big Data Analytics

The need for enterprise-class infrastructure is on the rise as more organizations evaluate and deploy big data technology. For instance, a 2013 Gartner study found that 64 percent of enterprises planned to invest in big data technology, up from 58 percent in 2012. Of that 64 percent, 30 percent have already made big data investments, 19 percent planned to invest within the next year, and 15 percent within two years.\(^1\)

Yet, investments are no guarantee for big data initiative success. In fact, a study by Infochimps, a big data cloud services provider, found that 55 percent of big data projects fail. Respondents to an Infochimps survey ranked the top five requirements for a big data platform as the ability to scale, ease of management, flexible architecture, speed to deployment, and security.\(^2\)

In many cases, those big data analytics requirements are not met by typical Hadoop sandbox environments which rely on commodity servers and DAS. IT professionals deploying commodity DAS based Hadoop face issues such as

- Limited interoperability with the overall IT infrastructure
- Increased data silos and duplicated “shadow IT” infrastructure
- Difficulty replicating data across applications and data centers
- Difficulties in implementing virtualization and multi-tenancy capabilities
- Lack of IT oversight and data visibility
- Security, data protection, and governance concerns
- Challenges in mission-critical performance, availability, and reliability
- Inflexibility in streamlining upgrades and migrating technology components
- High costs for additional scale-up hardware

Traditionally, big data practitioners have been attracted by the ability in a commodity server/DAS environment to readily add server nodes when a big data system requires additional resources. Yet as data volume requirements increase, IT often finds that it is buying excessive servers for more storage capacity because they cannot scale out storage separately from servers, as Hadoop requires three or more copies of data that need to be replicated.

The use of commodity servers with DAS has also been a popular practice for Hadoop deployments due to the notion that storage needs to sit close to the CPU and that nodes can easily be added for improved performance. However, this approach does not take into account the end-to-end data management processes of DAS. Time to deploy, ingest and manage data, and realize value increases significantly because of excessive movement of data from source systems, replication of data within the Hadoop cluster, and moving results for downstream data processing.

As an alternative to DAS, HDFS-enabled NAS enables a decoupling of compute and storage not possible in a server/DAS environment. The traditional gap in data transfer speed between server DAS and external storage is diminishing with advancements in interconnect networks and Ethernet connectivity speeds. Users can expect up to 100 GB per second of concurrent throughput on the Hadoop storage layer using EMC Isilon storage. In addition, for ultra-low latency processing, IT can use server-side flash to realize the protection, availability, and scale of the shared storage model while accelerating Hadoop job performance.

The need for reliable, available, and scalable enterprise-class storage is increasing as Hadoop-related storage infrastructure grows and the business value of timely analytics results increases. For example, an IDC study found that more than 60 percent of organizations surveyed were experiencing storage growth greater than 30 percent in their Hadoop environments.\(^3\)

True enterprise-class server technology coupled with best-in-class HDFS enabled

---

1 Gartner, “Gartner Survey Reveals That 64 Percent of Organizations Have Invested or Plan to Invest in Big Data in 2013,” news release, September 23, 2013.
NAS systems offers a clear path forward towards capitalizing on the business value of big data with outstanding IT control and cost-efficiency.

Why Vblock® Systems with EMC Isilon for Hadoop?

Converged infrastructure systems are a natural fit to enable organizations to evolve early big data implementations into a production environment that interoperates with the overall IT infrastructure. VCE has seen increasing demand and interest among enterprises interested in utilizing Vblock Systems, the world’s most advanced converged infrastructure, for their big data analytics solutions.

 Seamlessly integrating best-in-class compute and network technology from Cisco, storage technologies and data protection from EMC, and server virtualization and virtualization management from VMware, Vblock Systems represent the next evolution of IT—one that unleashes simplicity by delivering the extraordinary efficiency and business agility of virtualization and cloud computing.

VCE Vblock Systems are delivering quantifiable benefits to customer organizations in data centers. In a study of 11 VCE customers using Vblock Systems for data center operations, IDC calculated a return on investment (ROI) of 294 percent over 11.6 months.¹ IDC found the Vblock Systems customers realized four times faster deployment for new infrastructure, with 79 percent less staff effort and five times faster deployment for new services. Downtime was reduced by 96 percent and annual data center costs by 50 percent.

“...Vblock System implementations sped deployments, simplified operations, improved business-support agility, saved money, and freed staff to launch new applications, extend services, and improve user/customer satisfaction.”

²-³ IDC

Pre-integrated, tested, and validated as a VCE™ technology extension for EMC® Isilon® storage, EMC Isilon can be quickly deployed, easily distributed, and counted on to increase the predictability and timeliness of IT projects to extend the VCE value proposition for big data analytics. This solution is open for organizations to deploy Hadoop distributions and other analytics tools of preference such as SAS and Solunk and to support traditional databases and other systems and applications as part of the analytic value chain.

It also empowers businesses to virtualize Hadoop deployment with VMware vSphere Big Data Extension (BDE) for operational simplicity, better resource utilization with multi-tenancy, and enterprise-class scaling and availability. The combination of a Vblock System and EMC Isilon offers distinct advantages not available in any other solution, including

- **Elastic scale-out capacity and performance.** In conjunction with Vblock Systems, EMC Isilon supports a scale-out infrastructure for big data analytics that avoids the limitations of traditional approaches, making possible greater cost-effectiveness, data protection, data access, resilience and availability, and simplicity of management. The scale-out approach helps enterprises embrace new data and applications on an open, converged platform.

- **In-place unstructured data analytics.** EMC Isilon provides an analytic and storage environment for unstructured big data, for example, social media, web content, mobile data, files, multimedia, and machine-generated information from sensors and devices. The unstructured data and analytics support in EMC Isilon augments the structured data storage capabilities built into Vblock Systems through EMC Symmetrix VMAX, EMC VNX, and/or EMC XtremIO.

- **Availability and reliability at petabyte scale.** With its pre-engineered system, Vblock Systems can reduce downtime by up to 96 percent. Once data sets grow into the hundreds of terabytes and the petabyte realm, new availability, management, and scalability challenges arise, as there will always be one or more components in a degraded state at any point in time within the storage infrastructure.

---

As such, guarding against single points of failure and bottlenecks becomes a critical and highly complex issue. The combination of Vblock Systems and Isilon are ideally suited to address this enterprise-class requirement.

- **Built-in data protection.** Vblock Systems can be protected by Vblock® Data Protection, a family of solutions including EMC Avamar, EMC Data Domain, EMC RecoverPoint, and EMC VPLEX. To extend this integrated Vblock Systems data protection capability, EMC Isilon further safeguards data with a variety of techniques, including data encryption at rest (DARE) journaling, snapshotting and remote replication, and NDMP backups.

For more detailed information on high availability, reliability, and data protection, please see the EMC white paper, "High Availability and Data Protection with EMC Isilon Scale-Out NAS."

With a new approach based on a shared infrastructure model enabled by Vblock Systems with Isilon, your enterprise can easily develop and maintain specific services to run analytics, Hadoop processing, and other applications leveraging the data in place, without having to move and replicate across multiple servers and storage systems. Furthermore, organizations are maturing to a point where they are using enterprise class platform as a service (PaaS) to further scale and simplify application development in conjunction with big data and analytics projects. As shown in Figure 1, in advancing from proof-of-concept environments to production-ready infrastructures, organizations are able to

- Increase time to insights with big data and platform as a service (PaaS)
- Drive elastic, on-demand service enablement
- Run IT as a service with mission-critical availability and performance
- Meet security, governance, and compliance requirements
- Ensure high availability and performance
- Drive standardization across data datacenters while providing flexibility for change
- Adapt to deployment demand incrementally with mix and match of technology components

*Figure 1 illustrates key concepts and benefits of converged infrastructure for big data analytics with Hadoop.*
Vblock Systems: A Foundation for Unstructured Data Value

Vblock Systems serve as a foundation for deriving business value from big data with a pre-integrated, pre-validated converged infrastructure that can be quickly deployed, easily distributed, and counted on to increase the predictability and timeliness of big data projects.

Figure 2 illustrates the core technological components in Vblock Systems.

Uniquely designed to help organizations achieve the highest levels of reliability, security, availability, and performance, Vblock Systems works with EMC Isilon for unstructured data, and EMC Symmetrix VMAX, VNX and/or XtremIO storage systems for structured data. Vblock Systems provide proven capabilities for data protection and mobility, security and compliance, and intelligent systems management.

- **Data protection and mobility.** Vblock Systems contribute to true peace of mind by providing a choice of safeguards, optimized for virtualized or cloud environments. Highly reliable backup and recovery, data replication, business continuity, and workload mobility helps ensure critical protection for applications and data running on Vblock Systems.

- **Security and compliance.** Vblock Systems are engineered, architected, and hardened according to best practices for each component and enterprise-grade business objectives—to ensure the highest security and simplify often-burdensome compliance requirements.

- **Intelligent systems management.** VCE Vision™ Intelligent Operations provides centralized systems management, optimizing services for a converged infrastructure and integrating directly into VMware technologies. VCE Vision™ software also provides an extensible API for leading management tools to protect investments in existing management frameworks.
When deploying big data and analytics on Vblock Systems, businesses can benefit from standardizing on the optimized infrastructures across data centers while choosing the right set of compute, network, and storage for big data and analytic use cases. For example

- **Achieve elasticity, scalability, and reliability** on Vblock Systems utilizing Cisco servers and best-in-class networks, and EMC Isilon scale-out storage.

- **Ensure mission-critical readiness** of Hadoop deployments with EMC Isilon deployed on Vblock System family to exploit multi-tenancy, virtual/bare-metal implementation, HDFS integration, privacy, security, governance, and reliability advantage.

- **Drive operational and performance gains** running high-performance databases in conjunction with sophisticated Hadoop or unstructured analytics in an integrated environment.

- **Ensure faster, better business outcomes** by analyzing massive quantities of data and scale in the private and public cloud in a hybrid cloud model.

- **Ensure data protection** with VCE Data Protection, including EMC Avamar, EMC Data Domain, EMC VPLEX, and EMC RecoverPoint.

- **Future-proof big data analytic deployments** with ease of upgrade for compute, network, and storage and analytics tools in response to changing demand and availability of new technologies.

Table 1 summarizes solution highlights of big data analytics on Vblock Systems.

<table>
<thead>
<tr>
<th>Solution highlights</th>
<th>Big data analytics on Vblock Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Availability / Reliability</td>
<td>Enterprise- and service provider-class, 6 x 9’s</td>
</tr>
<tr>
<td>Upgradability / Scalability</td>
<td>Upgrade / move to / add infrastructure resources with data in place</td>
</tr>
<tr>
<td>Performance / Scalability</td>
<td>Enterprise- and service provider-class, high configurability with dynamic and intelligent scaling</td>
</tr>
<tr>
<td>Security / Privacy</td>
<td>Data-at-rest encryption, secure data transmission, multi-level segregation, control, and isolation</td>
</tr>
<tr>
<td>Protocol Support</td>
<td>HDFS 1/2, NFS, CIFS, FTP, HTTP</td>
</tr>
<tr>
<td>Data/Analytic Support</td>
<td>All data including structured and unstructured data with EMC Symmetrix VMAX, VNX, XtremIO, and/or Isilon</td>
</tr>
<tr>
<td>Multi-Tenancy</td>
<td>Yes</td>
</tr>
<tr>
<td>Mixed Workload</td>
<td>Yes</td>
</tr>
<tr>
<td>Integrated Support</td>
<td>Yes</td>
</tr>
<tr>
<td>Virtualization</td>
<td>Yes, as well as bare metal, implement VMware vSphere BDE</td>
</tr>
<tr>
<td>Operations Management</td>
<td>Yes with VCE Vision software, Open API. System metrics available</td>
</tr>
</tbody>
</table>
Business Benefits of Big Data Analytics on Vblock Systems and the VCE technology Extension for Isilon

Big data analytics on Vblock Systems enable your organization to achieve a complete view of the business, accelerating time to value from your infrastructure and Hadoop investments. The solution enables your IT team to deploy data and applications to meet business demands from users, while maximizing the flexibility and utilization of IT resources in a service-based model.

Table 2 summarizes key business benefits of Vblock Systems.

<table>
<thead>
<tr>
<th>Benefit</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed: Time to value</td>
<td>• Get up and running with 40+ day lead time ready to deploy and migrate big data and non-big data</td>
</tr>
<tr>
<td></td>
<td>• Meet 5x faster time for new services to traditional systems</td>
</tr>
<tr>
<td>Agility: Operational efficiency</td>
<td>• Elastically Scale storage/network/compute tuned to demand</td>
</tr>
<tr>
<td></td>
<td>• Rapidly deploy mixed workloads on a single platform to support more innovation—analyze any data regardless of variety, volume, or velocity</td>
</tr>
<tr>
<td></td>
<td>• Virtualized for multi-tenancy and optimized efficiency</td>
</tr>
<tr>
<td>Open: Standards-based with open source support</td>
<td>• Open and standards-based aggregation of functional, environmental, and operational characteristics</td>
</tr>
<tr>
<td></td>
<td>• Easy to incorporate existing and new big data and analytic workloads</td>
</tr>
<tr>
<td></td>
<td>• Support existing and emerging management and analytic frameworks</td>
</tr>
<tr>
<td>Scale: Enterprise- and service provider-grade</td>
<td>• Enterprise- and service provider-grade, suited to IT-as-a-service model</td>
</tr>
<tr>
<td></td>
<td>• Start for today’s demand, and adapt with future demand</td>
</tr>
<tr>
<td>Automation: Pre-validated turnkey platform</td>
<td>• Automated validation and testing and roadmap alignment with Cisco, EMC, and VMware.</td>
</tr>
<tr>
<td></td>
<td>• Reduce error-prone processes of deploying and maintaining IT infrastructure and applications</td>
</tr>
<tr>
<td>Predictability: Reduced cost, risk, and complexity</td>
<td>• Reduce down time by up to 96 percent</td>
</tr>
<tr>
<td></td>
<td>• Decrease annual data center costs by up to 50 percent</td>
</tr>
<tr>
<td></td>
<td>• Integrated support</td>
</tr>
</tbody>
</table>


Vblock Systems and Technology Extension Overview

Vblock Systems from VCE represents the next evolution of IT—one that unleashes simplicity by delivering the extraordinary efficiency and business agility of virtualization and cloud computing. Seamlessly integrating best-in-class compute, network, and storage technologies from industry leaders Cisco, EMC, and VMware, Vblock Systems provide dynamic pools of resources that can be intelligently provisioned and managed to address changing demands and business opportunities.

The expanded Vblock Systems portfolio is built on a single, highly secure, standardized, market-leading infrastructure. Pre-integrated, tested, and validated, Vblock Systems can be quickly deployed, easily distributed, and counted on to increase the predictability and timeliness of IT projects for solution providers and end-users alike. The portfolio is designed for organizations to incorporate the big data and analytic workloads including Hadoop utilizing the existing environment. The businesses can select the right systems...
to suit their workloads and their operating requirements from the portfolio and incrementally add resources and capabilities to support their data center journeys.

**VBLOCK SYSTEM 700 FAMILY**
Reliably runs thousands of virtual machines and desktops supporting mission-critical applications on SAP, Oracle, Microsoft Exchange, Microsoft SharePoint, VDI and more. The unprecedented scalability, configurability, high availability, systems visibility, and data protection make the Vblock System 700 family a perfect choice.

**VBLOCK SYSTEM 500 FAMILY**
The world’s first all-flash converged infrastructure system is ideal for applications that demand the highest throughput at the lowest latency, such as online transaction processing and analytics. Its performance-optimized data center footprint lets you consolidate multiple workloads without compromising performance or availability—for maximum IOPS and the lowest TCO.

**VBLOCK SYSTEM 300 FAMILY**
Enables the substantial scale needed for large virtualization and cloud implementations. Vblock System 300 is built to support mission-critical enterprise applications with the capacity, flexibility, and performance to run vast VDI environments, mixed workloads, and essential cloud services—while ensuring operations and management simplicity.

**VBLOCK SYSTEM 200 FAMILY**
Offers mid-sized organizations a highly efficient virtualized infrastructure to run their entire business. Tap the impressive power and capacity of the Vblock System 200 family in a broad range of configurations, with plenty of room for expansion. Supports a variety of applications—from core IT services to virtual desktop infrastructure (VDI).

**VBLOCK SYSTEM 100 FAMILY**
Provides the world’s most advanced converged infrastructure, scaled to address the needs of remote and distributed operations as well as mid-sized businesses. Choose from two Vblock System 100 models that you can deploy anywhere from a data center to a computer closet, and have operational in as little as 30 days from your order.

**VCE TECHNOLOGY EXTENSIONS**
VCE technology extensions provide an innovative way to expand and enhance Vblock Systems converged infrastructure. The technology extensions let you quickly and dramatically increase processing power or add storage capacity, without typical technology risks.

**EMC Isilon Scale-Out NAS for In-Place Hadoop Analytics**

EMC Isilon transforms data analytics by coupling a key tool of data science, Hadoop, with the natural home of big unstructured data—scale-out NAS. By its very nature, unstructured data flows into large storage systems, typically via SMB/CIFS and NFS NAS protocols. Meanwhile, data scientists and other data analysts are increasingly turning to Hadoop to analyze unstructured data.

By allowing Hadoop clients direct access through Hadoop Distributed File System (HDFS) to the data that is already stored in an Isilon cluster with NFS, HTTP, FTP, or SMB, an EMC Isilon cluster pairs the standard tool of data science with a highly scalable storage system. Combining Hadoop with scale-out NAS supports a fundamental shift taking place in advanced organizations: Businesses are trying to analyze their data and extract value from it in an efficient and timely manner. Isilon fosters data analytics without placing a heavy reliance on Hadoop infrastructure resources and without ingesting and/or migrating data into stand-alone HDFS solution.

An EMC Isilon cluster delivers value by separating data from compute and streaming data ingest for achieving optimized balance of compute and storage. With an EMC Isilon cluster, you can store data on an enterprise storage platform with your existing workflows and standard protocols, including SMB, HTTP, FTP, Representational State Transfer (REST), and NFS. Regardless of whether you store the data with SMB or NFS,
you can analyze it with a Hadoop compute grid through HDFS. There is no need to set up a separate HDFS cluster and then move data to it with tedious HDFS copy commands or specialized Hadoop connectors. Instead of struggling to take your data to Hadoop, you can bring Hadoop to your data.

In concert with Vblock Systems, EMC Isilon helps you adapt to fluid storage requirements, non-disruptively add capacity and performance in cost-effective increments, reduce storage overhead, and exploit your data through in-place analytics.

Table 3 highlights distinctions between traditional Hadoop storage and EMC Isilon.

<table>
<thead>
<tr>
<th>Traditional Hadoop Storage Implementation</th>
<th>EMC Isilon Storage Implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dedicated Storage Infrastructure</td>
<td>Scale-Out Storage Platform</td>
</tr>
<tr>
<td>• One-off for Hadoop only</td>
<td>• Multiple applications and workflows</td>
</tr>
<tr>
<td>Downtime</td>
<td>No Downtime</td>
</tr>
<tr>
<td>• Failover process often causes downtime</td>
<td>• Distributed NameNode</td>
</tr>
<tr>
<td>Lacking Enterprise Data Protection</td>
<td>End-to-End Protection</td>
</tr>
<tr>
<td>• No Snapshots, replication, backup</td>
<td>• SnapshotIQ, SyncIQ, NDMP Backup, Data at Rest Encryption (DARE), Rose based security, File-based ACL</td>
</tr>
<tr>
<td>Poor Storage Efficiency</td>
<td>Industry-Leading Storage Efficiency</td>
</tr>
<tr>
<td>• 3x mirroring</td>
<td>• 80% Storage Utilization</td>
</tr>
<tr>
<td>Fixed Scalability</td>
<td>Independent Scalability</td>
</tr>
<tr>
<td>• Ridged compute to storage ratio</td>
<td>• Add compute and storage separately</td>
</tr>
<tr>
<td>Manual, Slow Import/Export</td>
<td>Multi-Protocol</td>
</tr>
<tr>
<td>• No protocol</td>
<td>• Industry standard protocols</td>
</tr>
<tr>
<td></td>
<td>• NFS, CIFS, FTP, HTTP, HDFS</td>
</tr>
</tbody>
</table>

An Isilon cluster simplifies data management while cost-effectively maximizing the value of data, distinguished by scale-out storage, separation of data from compute, native Hadoop integration, robust data protection, high resilience and availability of enterprise-class security and compliance, and multi-protocol support.

Scale-Out Storage to Meet Big Data Demands

Traditional DAS are not up to the big data challenge in a highly scalable and efficient way as they require the use of either multiple copies of data or many disparate volumes, volume management, underlying RAID and defragmentation that drives up overhead. As big data and business demands increase, IT faces challenges with scale-up systems that depend on a single controller to manage storage.

In a white paper on managing big data growth, IDC argued that the limitations of scale-up storage and file servers require “deployment of new classes of storage solutions (scale-out storage solutions) that are optimized for rapid data ingest, efficient storage management, and reliable access. Enabling a distributed scale-out architecture; particularly as it relates to data storage, protection, and access; is the foundational building block to being able to evolve organizations from application-focused to data/information-focused enterprises.”

EMC Isilon addresses these storage needs, while also provides an alternative file system to HDFS, which lacks certain resiliency and robust data management capabilities, such as lack of tiering and single point of failure within the NameNode. “File systems like… EMC’s Isilon OneFS that have earned a reputation for their robust scale-out capabilities are clearly preferred as alternatives to HDFS,” IDC said in a report on Hadoop deployments.

---


Isilon scale-out NAS is a fully distributed system that consists of modular hardware nodes arranged in a cluster. The distributed EMC Isilon OneFS operating system combines the nodes’ memory, I/O, CPUs, and disks into a cohesive storage unit that presents a global namespace as a single file system. The nodes work together as peers in a shared-nothing hardware infrastructure with no single point of failure. Each node adds capacity, performance, and resiliency to the cluster, and each node acts as a Hadoop NameNode and DataNode. The NameNode daemon is a distributed process that runs on all nodes in the cluster, making it implicitly HA.

As nodes are added, the file system expands dynamically and redistributes data, eliminating the work of partitioning disks, creating volumes and rebalancing data. The result is a highly efficient and resilient storage architecture that delivers all the advantages of scale-out NAS for big data analytics with Hadoop shown in Figure 3.

“Scale-out storage systems have proven to be a perfect remedy for coping with the flood of unstructured data inundating enterprise IT.”

—Kurt Marko, InformationWeek

Figure 3 illustrates Isilon scale-out NAS (storage layer) connected to Hadoop (compute layer).

Separation of Data from Compute
EMC Isilon scale-out NAS fosters the convergence of Hadoop data analytics with stored data. Instead of requiring Hadoop users to move data to the Hadoop cluster, administrators can point the Hadoop compute function to data residing on EMC Isilon, streamlining the entire analytics workflow. This convergence eliminates the need to extract data from a storage system and load it into a traditional Hadoop environment, and then to export the data and results after analysis.

Streamlining the analytic workflow speeds the transition to a data-driven enterprise with faster, easier, and more flexible analytics and without heavy reliance on IT. Business personnel can use existing workflows, especially via SMB and NFS protocols, to collect and manage data they want to analyze. For example, if technical support staff has stored log data from a year’s worth of support cases on a file server, support managers can analyze data in place to identify problems and areas for improvement.

---

Decoupling the Hadoop compute and HDFS storage through EMC Isilon shared storage can provide equivalent performance and improved time to results, with lower CAPEX and OPEX costs in many cases. With advancements in interconnect networks and in particular Ethernet connectivity speeds, the traditional gap in processing speed between server DAS and external storage is shrinking rapidly. In addition, for ultra-low latency processing, organizations can add server-side flash like EMC XtremSF to create a hybrid architecture for the protection, availability, and scale of the shared storage model while still selectively performing low latency data processing with enterprise servers powered by server-side flash.

Native Hadoop Integration with Distribution Portability

EMC Isilon is the first and only scale-out NAS storage platform that integrates native HDFS as a protocol to help enterprises streamline deployment and bring enterprise-proven features to Hadoop workflows. In addition, EMC Isilon fully supports Apache Hadoop and all leading Hadoop distributions, including Apache Hadoop, Pivotal HD, Cloudera, and Hortonworks with portability between distributions. If you start off with one distribution and later realize it is not supporting your needs, you simply switch to another distribution without migrating any data. In fact, you can simultaneously run several different Hadoop distributions against the same data or different data managed in EMC Isilon.

In a Hadoop/Isilon environment, Hadoop clients running MapReduce jobs access data stored on an Isilon cluster over HDFS. The OneFS operating system becomes the native HDFS storage for MapReduce clients and implements the HDFS protocol on every node on chosen compute nodes, with each Isilon node functioning as both a NameNode and DataNode. An Isilon node, however, does not act as a job tracker or task tracker; those functions remain the purview of Hadoop compute clients. Because each Isilon node functions as a NameNode, the function of a Secondary NameNode—which checking the internal NameNode transaction log—is unnecessary.

The Isilon cluster automatically load-balances HDFS connections across all nodes in the cluster. Because OneFS stripes Hadoop data across the cluster and protects it with parity blocks at the file level, any node can simultaneously serve DataNode traffic as well as NameNode requests for file blocks. A virtual racking feature mimics data locality. For example, you can create a virtual rack of nodes to assign compute clients to the nodes closest to a client’s network switch to interoperate with existing network topology or optimize performance.

Robust and Efficient Data Protection to Drive Lower Maintenance

Before data analysis can begin in traditional Hadoop environments, the entire data set must be ingested, and three or more copies of each file system block created as a means of data protection against hardware failure. Once analysis is complete, the results need to be exported. These costly and time-consuming processes are also prone to potential security and compliance exposure.

For example, if you wanted to store 4 PB of Hadoop data, you would need more than 12 PB of raw disk capacity in a traditional Hadoop cluster using a default 3x mirroring to store data in it. Storing the same 4 PB of data in Isilon for Hadoop with OneFS, however, requires only 5 PB of raw disk capacity in the Isilon cluster, as reflected in Table 4. This results in a significant CAPEX savings as well as a much simpler environment to manage.

<table>
<thead>
<tr>
<th>File System</th>
<th>Capacity Goal</th>
<th>Consumption with overhead</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDFS</td>
<td>4 PB</td>
<td>12 PB of disk space</td>
<td>4 PB plus 3 copies of each block</td>
</tr>
<tr>
<td>OneFS</td>
<td>4 PB</td>
<td>5 PB of disk space</td>
<td>4 PB plus FEC protection</td>
</tr>
</tbody>
</table>
With the Isilon data protection scheme, more than 80 percent of an Isilon cluster’s capacity can be utilized, bringing efficiency to a data analytics workflow. In contrast to HDFS and its triple replication for every block, the efficiency of Isilon data protection optimizes ROI and reduces total cost of ownership (TCO).

Instead of replicating the data as is done in HDFS, OneFS stripes the data across the cluster over its internal InfiniBand network and protects the data with forward error correction (FEC) codes. FEC is a highly efficient method of reliably protecting data. FEC encodes a file’s data in a distributed set of symbols, adding space-efficient redundancy. With only a part of the symbol set, OneFS can recover the original file data. In a cluster with five or more nodes, FEC delivers as much as 80 percent efficiency. As you add nodes to a cluster, data protection becomes increasingly efficient.

Striping data with FEC codes consumes much less storage space than replicating data three times—as much as 2.5 times fewer drives. Stripping data lets a Hadoop client connecting to any node take advantage of the entire cluster’s performance to read or write data. If you set the replication level from an HDFS client, OneFS ignores it and instead uses the protection level that you set for the directory or the file pool that contains your Hadoop data.

By default, OneFS optimizes striping for concurrent access. With Hadoop, however, the dominant data access pattern might be streaming. You can set OneFS to lay out data for streaming access patterns to increase sequential read performance for MapReduce jobs. To better handle streaming access, OneFS stripes data across more drives; additionally, OneFS utilizes advanced caching and prefetch algorithms to optimize concurrent streaming performance. Streaming is most effective on directories or subpools serving large files or handling large compute jobs.

High Resilience and Availability

The Isilon architecture provides a resilient foundation for data availability, with no single point of failure, tolerance for multi-failure scenarios, NameNode redundancy, DataNode load balancing, a fully distributed single file system, and unparalleled levels of data protection. In a 2013 report, Gartner rated EMC Isilon highest among storage vendors for resilience, a platform’s capabilities for provisioning a high level of system availability and uptime. In a traditional Hadoop cluster using DAS, a single NameNode manages file system namespace operations. In an Isilon environment for Hadoop with OneFS, every Isilon node can reply to NameNode or DataNode requests shown in Figure 4. This allows for I/O load balancing from multiple mapper and reducer tasks that run on multiple Hadoop compute nodes to occur across multiple Isilon nodes. In this way, OneFS also eliminates single points of failure that exist in a traditional Hadoop cluster.

Figure 4 illustrates how Isilon OneFS runs a daemon that speaks NameNode and DataNode natively.

---

In a traditional DAS Hadoop environment, a disk drive failure will cause the Hadoop jobs utilizing that data to timeout, fail and be restarted on a different copy of data. At the same time, the data that was on the failed drive will be re-replicated over the same network that Hadoop jobs are utilizing—this leads to extended time to results and unpredictable SLAs. With Isilon, a drive failure will not cause any job timeouts, and will have minimal impact on SLAs, since data reprotection is done over Isilon’s private Infiniband network.

The Isilon SmartConnect feature contributes to data high availability by supporting dynamic failover and failback for Hadoop compute clients. This ensures that if a node or a network link fails, all in-flight reads and writes associated with a MapReduce job are handed off to another node in the Isilon cluster with no interruption to the MapReduce job.

**Enterprise-Class Security and Compliance**

EMC Isilon solutions offer robust security options, including role-based access control (RBAC) to provide a clear separation between storage administrators and file system access, and SEC 17a-4 compliance write once/read many (WORM) data protection to help prevent accidental or malicious data deletion or alteration. EMC Isilon solutions allow you to establish authentication zones that serve as secure, isolated storage pools for specific departments within your organization. EMC also offers file system auditing to help organizations address important regulatory requirements such as FISMA, HIPAA, and Sarbanes-Oxley. Isilon Data at Rest Encryption provides FIPS-compliant data encryption, which protects against data accidentally or maliciously leaving the data center.

In addition, integrated support for the EMC Common Event Enabler (CEE) in OneFS provides access to audit software applications to allow the recording of all file access over SMB/CIFS protocols and to facilitate report creation and chain of custody tracking to meet regulations such as Sarbanes-Oxley and HIPAA. OneFS also strengthens security with capabilities to authenticate connections with the Kerberos protocol, and to provide granular controlled data access via file-level Access Controls.

**Multi-Protocol Support**

Isilon solutions include integrated support for a wide range of industry-standard protocols, including NFS, SMB (or 1/2 CIFS), HTTP, HFDS, FTP and REST-based object access for your cloud initiatives. This capability combines HDFS with posix-compliant protocols. As a result, your IT personnel spend less time importing and exporting data into a cluster that only supports HDFS. Isilon multi-protocol support promotes a highly flexible storage infrastructure that allows you to eliminate data silos, simplify workflows, accelerate business analytics, support cloud infrastructure analytics, and derive greater value from your enterprise applications and data.

**Big Data and Analytic Use Case Examples**

Organization in a broad range of industries, including market intelligence, technology, telecommunications, pharmaceuticals, financial services, retail, food and hospitality, and others are exploring converged infrastructure as a key foundational element of their big data analytics initiatives.

**Market Intelligence Researcher**

A large market intelligence researcher in the fashion industry uses EMC Isilon scale-out NAS as a foundational element of an aggressive program to capture and analyze fashion products data from leading online stores. Along with the Pivotal Hadoop distribution and Pivotal Greenplum database, EMC Isilon enables the company to transform raw data into market intelligence to help retailers make smarter merchandising decisions.

The solution enabled the business to meet its objective for rapid launch of the offering going live in less than four weeks. Their big data analytics systems captures more than three million data points on products, pricing, and sales across 250,000 goods. EMC and Pivotal have delivered the performance and stability needed for the launch, with scalability to grow.
A Leading Managed Hosting Provider
A leading managed hosting provider with a range of infrastructure-as-a-service (IaaS) offerings, uses EMC Isilon as a Tier One target for rich streams of Hadoop data as part of a broader infrastructure that relies on VCE, EMC, and VMware technologies in delivering backup, recovery, archiving, and other hosted services. One customer of this hosting provider, a large hydro business in the Netherlands, uses the Isilon/Hadoop system for analysis and storage of large volumes of data on water height captured through sensors around the country. "We told them we had Isilon, which they can connect to with Windows Sharing or over HDFS, and that can be easily scaled to meet capacity needs. They are now using Isilon as their Tier One storage platform for all Hadoop data because of the reliability it provides.

Telecommunications Leader
A global telecommunications leader had a business imperative to use network traffic to create service offerings and capture updated information to monitor the business and compliance with service-level agreements (SLAs). The business needed to find ways to aggregate structured customer data with network data for soft alarming, anomaly detection, and prevention of outages. Their vision was to transition from centralized enterprise data warehouse (EDW) to edge analytics to increase agility and reduce latency. They also needed options to use both virtualized and bare-metal deployment. In transforming the environment, the company chose to modernize the environment to reduce latency and risk for data aggregation, parsing, computation, and movement. They achieved the business outcome that accelerates faster insight with a simplified, standardized approach to replicate across data centers.

Pharmaceutical Giant
A pharmaceutical giant needed to increase visibility into department-level use of Hadoop and better incorporate departments into enterprise IT across a standardized Hadoop environment for big data analytics. The initiative would minimize data silos and IT sprawl created by data scientists in multiple departments. The business driver was to modernize legacy environment in multiple business units that have varying objectives—streamline R&D processes, understand partner sales, and governmental approval. Having chosen Vblock Systems with EMC Isilon for their standardized platform, the company wanted to extend the environment and deploy analytic and other related applications with center of excellence in a service based model. With the introduction of the new solution, the company was able to reduce the amount of wait time for service provisioning and increased utilization of systems in supporting business needs in an accelerated fashion.

Global Financial Services Provider
A global financial services provider had many Hadoop projects and encountered difficulties in deploying Hadoop on commodity servers in extended production environments. Hadoop on commodity servers was not interoperable with the rest of the data value chain, and introduced issues with excessive downtime and the risk of losing client data. The company is seeking a more reliable hardware infrastructure with better posture for compliance, governance, and security. This drove them to create a new operational structure to handle Hadoop on Vblock Systems with EMC Isilon in a reliable, protected environment to strengthen analytic practices. The solution was created as an integrated system that can handle analytic and non-analytic workloads and ensure data protection back-up while ensuring co-existence with current information management processes.

Multi-Channel Retail Chain
A large retailer, known for its analytic capabilities, had difficulties getting infrastructure systems for big data analytics with Hadoop ready up and running quickly. They were looking for an alternative in converged infrastructure to accelerate go-live while taking advantage of structured and unstructured data. The retailer needed a better, open platform to handle growth in the variety, volume, and velocity of data and a large number of analytic and transactional applications. The business had difficulties in closing the analytic process gap across a multi-channel supply chain to demand at stores and web. IT chose the integrated Hadoop environment with point-of-sale logs for better analytics and decision-making. They were able to extend the system to other countries not only to accelerate time to insight but to reduce cost and risk by standardizing globally.
Major Food Chain
A major fast food chain had a corporate mandate to profile customer behaviors and offer choices that would help retain loyalty and increase share-of-wallet per customer. They had challenges optimizing store performance based on unique characteristics. They used to run a multitude of analytics separately, but needed to take a unified approach to fully understand customer-centric approach of timing, serving, approaching, and offering what customers want or may want in the future. Their goal was to analyze performance across 35,000 stores—customer interactions, in-store traffic, drive-through, staffing, layout, menus, and orders. The solution was designed to examine videos, sensors, POS data, and other data to drive analytics and decision-making. They were able to strengthen the analytics and decision-making capabilities to support better trade promotions, merchandising optimization, and customer marketing.

Vblock Systems with VCE™ technology extension with EMC® Isilon®
Vblock Systems are an ideal foundation to take advantage of EMC Isilon for increased performance and time to results for file-based data applications and analytic workflows—all from a single file system architecture. You can scale both capacity and performance immediately to meet your specific business needs without additional IT burden. There is no need to over-provision storage capacity or performance: Simply purchase storage when you need it and automatically scale it. Additionally, you can create analytics solutions which leverage standard open-source Hadoop software, but provide enterprise data governance and management capabilities, and bring the analytics tools to the data, rather than create a standalone siloed analytics infrastructure.

By combining Vblock Systems and Isilon scale-out architecture with vSphere Big Data Extension as shown in Figure 4, organizations can take advantage of extending the existing Vblock Systems environment or developing a new environment with Isilon. It offers adaptability, protection and scalability from compute, storage and network within a Vblock System, as well as a scale-out Isilon solution with native Hadoop integration capability.

Figure 4 illustrates Schematic on how Vblock Systems and Isilon can be used together for virtualized Hadoop.
Conclusion

Business and IT face growing pressures and challenges to fine-tune big data investments and blend big data analytics into the broader enterprise ecosystem to drive greater business value. Time to results, performance, scalability, reliability, security, and reduced TCO are all of paramount concern as the big data era continues to unfold. The combination of Vblock Systems and EMC Isilon scale-out NAS for in-place Hadoop analytics is increasingly the solution of choice for enterprises serious about unleashing business value from big data.

In this paper, we

- Introduced the solution benefits of Vblock Systems and the VCE technology extension for EMC Isilon storage
- Provided an overview of Vblock Systems and EMC Isilon extensible for big data and analytics
- Highlighted the challenges and limitations of commodity servers and DAS for Hadoop
- Outlined key functional capabilities of Vblock Systems and EMC Isilon

Organizations seeking to gain competitive advantage through big data analytics leveraging enterprise Hadoop can deploy Vblock Systems with EMC Isilon as an enterprise-grade solution to many of today’s most difficult big data challenges, with scalability and innovation for sustainable value in the years to come.

Next Steps

To learn more about this and other solutions, contact a VCE representative or visit www.vce.com.

ABOUT VCE

VCE, formed by Cisco and EMC with investments from VMware and Intel, accelerates the adoption of converged infrastructure and cloud-based computing models that dramatically reduce the cost of IT while improving time to market for our customers. VCE, through the Vblock Systems, delivers the industry's only fully integrated and fully virtualized cloud infrastructure system. VCE solutions are available through an extensive partner network, and cover horizontal applications, vertical industry offerings, and application development environments, allowing customers to focus on business innovation instead of integrating, validating, and managing IT infrastructure.

For more information, go to vce.com.