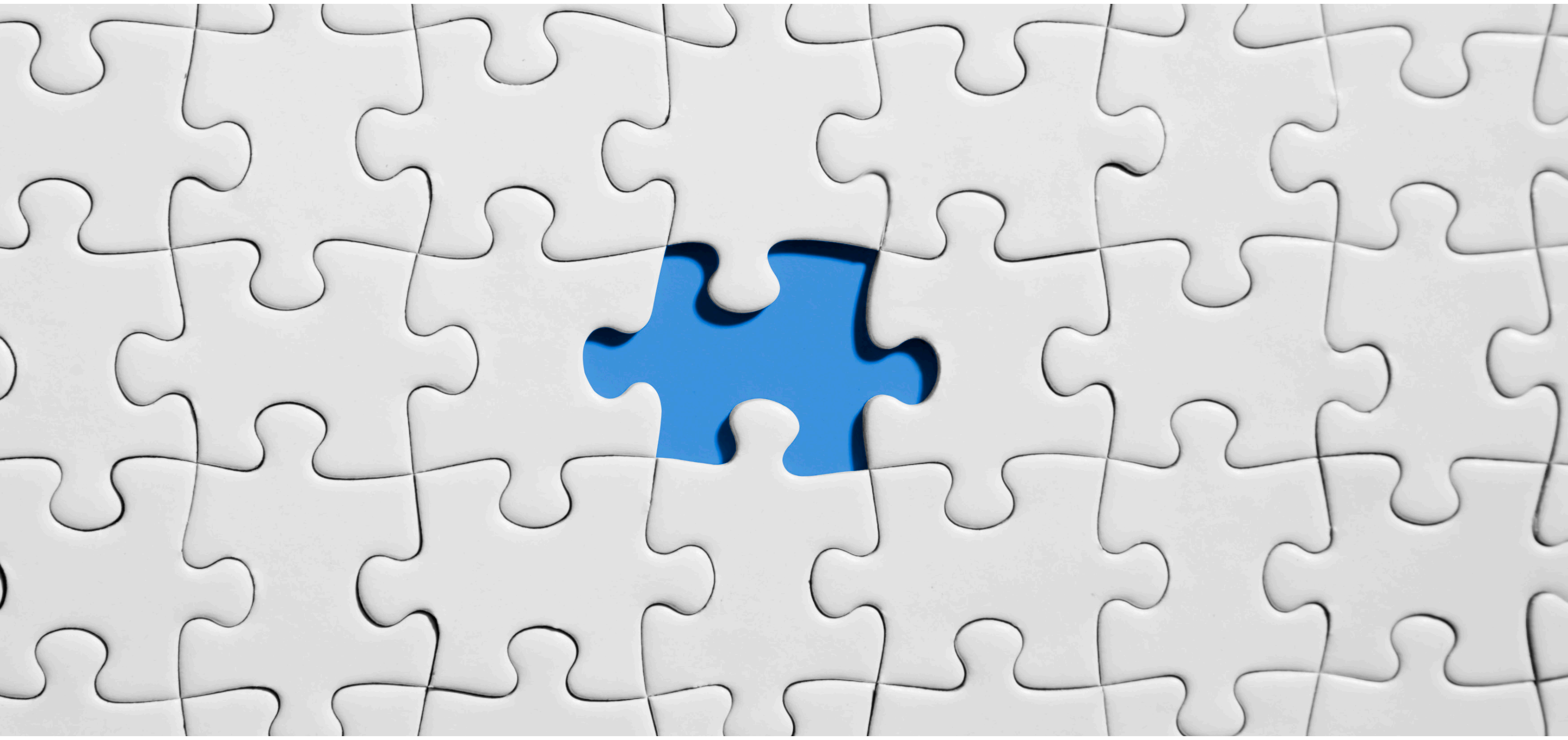
DELL Technologies

USE DELL TECHNOLOGIES TO SOLVE YOUR DARK DATA CHALLENGES



Sonali Singh
Sonali.s@dell.com

Knowledge Sharing Article

DELL Technologies Proven Professional

The Dell Technologies Proven Professional Certification program validates a wide range of skills and competencies across multiple technologies and products.

From Associate, entry-level courses to Expert-level, experience-based exams, all professionals in or looking to begin a career in IT benefit from industry-leading training and certification paths from one of the world's most trusted technology partners.

Proven Professional certifications include:

- Cloud
- Converged/Hyperconverged Infrastructure
- Data Protection
- Data Science
- Networking
- Security
- Servers
- Storage
- Enterprise Architect

Courses are offered to meet different learning styles and schedules, including self-paced On Demand, remote-based Virtual Instructor-Led and in-person Classrooms.

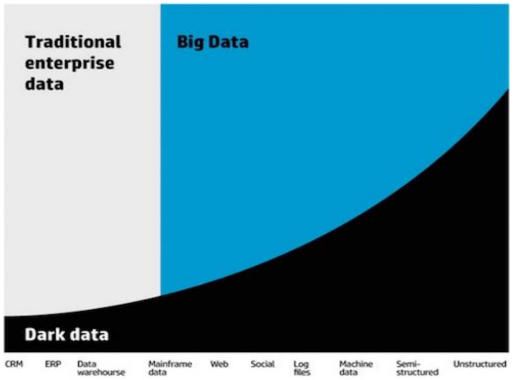
Whether you are an experienced IT professional or just getting started, Dell Technologies Proven Professional certifications are designed to clearly signal proficiency to colleagues and employers.

Learn more at www.dell.com/certification

Table of Contents

Disclaimer: The views, processes or methodologies published in this article are those of the author. They do not necessarily reflect Dell Technologies' views, processes or methodologies.

Introduction



Dark data is defined by Gartner as “the information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes (for example, analytics, business relationships and direct monetizing).” Or data that business and industry are paying to store, protect and manage yet isn’t being efficiently utilized to improve the value of their business. According to research by the Compliance, Governance and Oversight Counsel, 69% of a company’s stored data has no value to the organization. So, why retainin the unused and unmanaged data? Therefore, CGOC proposes use of content-based retention policies that will empower companies to keep only what is important to the business, hence gaining more benefits. According to Gartner, keeping only Business-valued data can reduce retention costs by up to 70 percent.

Types of data: Datum of data is now fueling the digital, cloud computing and Artificial intelligence era. Thus, it is important to know the data type we are dealing with. These are;

- **Big Data:** Large data sets comprised of both structured and unstructured data and leveraged strategically by business to inform decisions and drive future growth.
- **Structured, unstructured and semi-structured data:** Data organized in a defined pattern such as a table or matrix and is easy to use is known as “**Structured data”**. Data not in an organized format and requires advanced expertise and software tool to use it is referred as “**Unstructured data”**. Structured data not in organized format is called “**Semi-structured data”.**
- **Machine data:** Digital information generated by computer, Embedded system IoT and other network devices.
- **Spatiotemporal data:** Data generated considering time and space, i.e. information generated to GPS.
- **Dark data:** Unstructured, untagged and untapped data that reside in storage repositories and has not been analyzed or processed.
- **Real-time data:** Information delivered as generated immediately without time lag, i.e. data used for tracking or navigation purpose.

Approaches to uncover dark data

Managing the growth of storage: This approach ensures that the organization keeps only data that has value to the business by managing the storage through tiered hierarchy as per types of data, lifecycle management tools and advanced storage platforms such as Converged platform and virtualization in existing data centers.

Deliver self-service access to workforce: Each group of users in an organization require different type(s) of data to support their duties. Giving users self-service capability to search for the data they need from the organization's archives as per the requirement maximizes workforce productivity with transformational business insights into the stored data.

Automation of the data lifecycle: Establishing governance policies for defensible content deletion can significantly reduce "dark data". This can be achieved with lifecycle management (LCM) tools and Monitoring tools. These tools not only help to manage the data efficiently but also analyze it for future growth outcomes.

Dark data is usually not used by enterprises for decision making because of less bandwidth or technical expertise or low data value. All reasons are valid for organizations to ignore this data. However, with various advances in technology, and ability to source, ingest, store and analyze large volumes of data by correlating it with other data sources, it becomes important for organizations to recognize this largely untapped data.

As per an IDC report, by 2025 data will increase by 66 zettabytes out of which 90 percent of data will be untagged data. Due to immersive growth in data, enterprises have moved to ingesting the data volume into massive storage. It makes sense to store this data and tag it as it is being stored. Extracting metadata out of this data will be key to exploiting the data, which can be profiled and explored using many tools available including visualization products. Advances in Machine Learning and Cognitive Computing combined up storage and increased processing power, opening possibilities of leveraging "dark" data intelligently.

What is Big Data?

Big Data is the massive volume of unstructured and structured data that inundates businesses on daily basis. Big Data can be analyzed for insights that lead to better decisions and strategic business moves.

A major chunk of raw data is unstructured data, out of which only 10% is analyzed; the rest of unstructured data remains untouched. This raises the question of how Big Data and Dark Data are interrelated. While Dark Data is a subset of Big Data, the majority of Big Data is comprised of Dark Data only. In essence, analysis of Big Data in effect fulfills the analysis of Dark Data also.

Therefore, it is mandatory to apply analytics on untouched data (Big Data). Analysis of Dark Data using tools such as Hadoop or Splunk integrated with Big Data analytics is called Dark Data analysis.

Big Data analytics enables processing of large data sets that contain a variety of data types to discover hidden patterns or correlations that provide actionable insights for business advantage. In search of faster ways to process large amounts of data, customers are looking to new applications like Hadoop, an open source programming framework, Splunk, Kafka, Spark and many more that support processing of large data sets in a distributed compute environment.

Trends in digital transformation described earlier have led to an explosion of analytics techniques that include predictive analytics, IoT analytics, Advanced Machine Learning/Deep Learning, Real-time analytics, etc. This has led to a need to deploy infrastructure that runs this diverse set of software tools and supports the multitude of analytics techniques. A “one size fits all” approach to deploying analytics infrastructure, be it on-prem or in the cloud, is no longer an optimal approach.

Role of Dell Technologies in Dark Data

Dell Technologies provides an end-to-end portfolio of pre-designed, integrated and validated tools for Big Data Analytics.

Dell Technologies provides Hadoop-ready architecture designed to address data analytics requirements, reduce costs and deliver outstanding performance. They offer a wide range of products on which the Hadoop framework is deployed efficiently such as Isilon, ECS, and Hyperconverged VxRail.

Integrated Isilon and Hadoop

Dell EMC Isilon is network attached storage that provides direct access of Big Data to Hadoop clients through HDFS. Powered by OneFS operating system, Isilon delivers scalable pool storage with global name space.

One FS conglomerates memory I/O, CPU and disk of the node into cohesive storage and presents a global namespace as a single file system.

Flexibility
Out-of-box Enterprise-grade Capabilities
Extreme Scale
High Performance
Analytics Solution Accelerators

Isilon Delivers

- Extreme Scale
- Limitless Capacity
- Efficient Storage
- Massively Parallel Machine Learning

➢ **Flexibility of Analytics Apps/Process**

- Analytics in-place over consolidated data
- Enable DevOps Process
- Deliver Analytics-as-a-Service

➢ **High Performance Data Analytics**

- Low Latency
- High Throughput
- High IOPS

➢ **Out-of-box Enterprise-grade**

- Security
- Data Protection
- Data Management
- Policy & Regulatory Compliance

How HDFS works with Isilon

The Hadoop file system (HDFS) can support thousands of nodes with rapid data transfer between the nodes. It uses the Map Reduce Technique that computationally divides the workload which multiple device can Process. Thus, if you want to analyze data generated on a hyper-converged infrastructure, you would need to extract the data, translate and load it on to Hadoop.

In a Hadoop implementation on an Isilon cluster, Isilon OneFS serves as the file system for Hadoop compute clients. The Hadoop distributed file system (HDFS) is supported as a protocol, which is used by Hadoop compute clients to access data on the HDFS storage layer.

Hadoop with Isilon enables to perform Data analytics Unstructured data spread across direct attached storage or Isilon.

Figure 3 depicts the Isilon Cluster workflow.

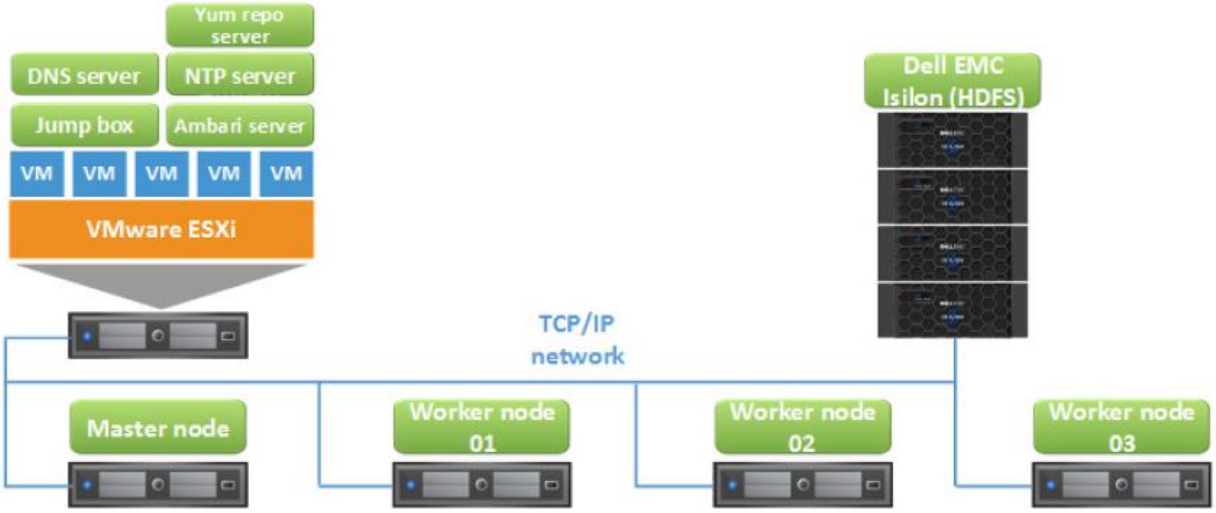


Figure 3. Hadoop tiered storage with Isilon solution architecture

- The Hadoop compute and Isilon storage layers are on separate clusters.
- Data is stored on Isilon cluster functioning as both Name Node and Data Node and compute layer is deployed on Hadoop cluster that is separate from Isilon cluster.
- HDFS protocol is implemented on OneFS between Isilon cluster and Hadoop Compute cluster. Hadoop Clients access the data on Isilon cluster over HDFS protocol.
- OneFS multiprotocol acceptance enables Hadoop clients to access data on Isilon cluster using NFS, SMB, FTP and HTTP protocol.
- Hadoop compute clients can connect to any node on the Isilon cluster that functions as a Name Node instead of being routed by a single Name Node.

Converged Platform

Dell Technologies offers ready-converged hardware architecture and software for Splunk.

VxRail is a preconfigured and pretested VMware hyper-converged infrastructure appliance. Powered by industry-leading VMware vSAN and vSphere software, the VxRail appliance streamlines and extends the VMware environment while dramatically simplifying IT operations with a known and proven building block for the software-defined data center.

Splunk

Splunk is an advanced, scalable, and effective technology that extracts log files stored in a system. It analyzes the machine-generated data to get operational intelligence. Splunk is an open platform application and does not require any database to store its data.

Major steps of Splunk are **searching, indexing and correlating** real time data. It searches, examines and correlates real time data generated from machine (mostly Big Data) and puts it in containers where further analysis of data takes place, after which graphs, metrics, and reports are generated as an outcome. It helps organizations take accurate decisions for future growth. For example, if system/machine facing issue, data generated by machine is in a scrambled format and cannot be understand by humans. Splunk helps the system administrator place that data into containers, perform algorithm and take out relevant data to locate the problem easily.

Why Splunk for Big Data Analytics

Splunk can collect and index any type of machine data. The following parameters make Splunk go-to software for Big Data.

- **Easy to Deploy and Use**

  Connect to your data with just a few clicks and easily create powerful dashboards.

- **Real-Time Alerts**

  Monitor your data to spot trends and identify specific patterns of activity or behavior.

- **Massively Scalable**

  Easily scale from a single server to multiple data centers.

- **Robust Security**

  Secure data handling, role-based access controls, auditability and data integrity.

Splunk With VxRail

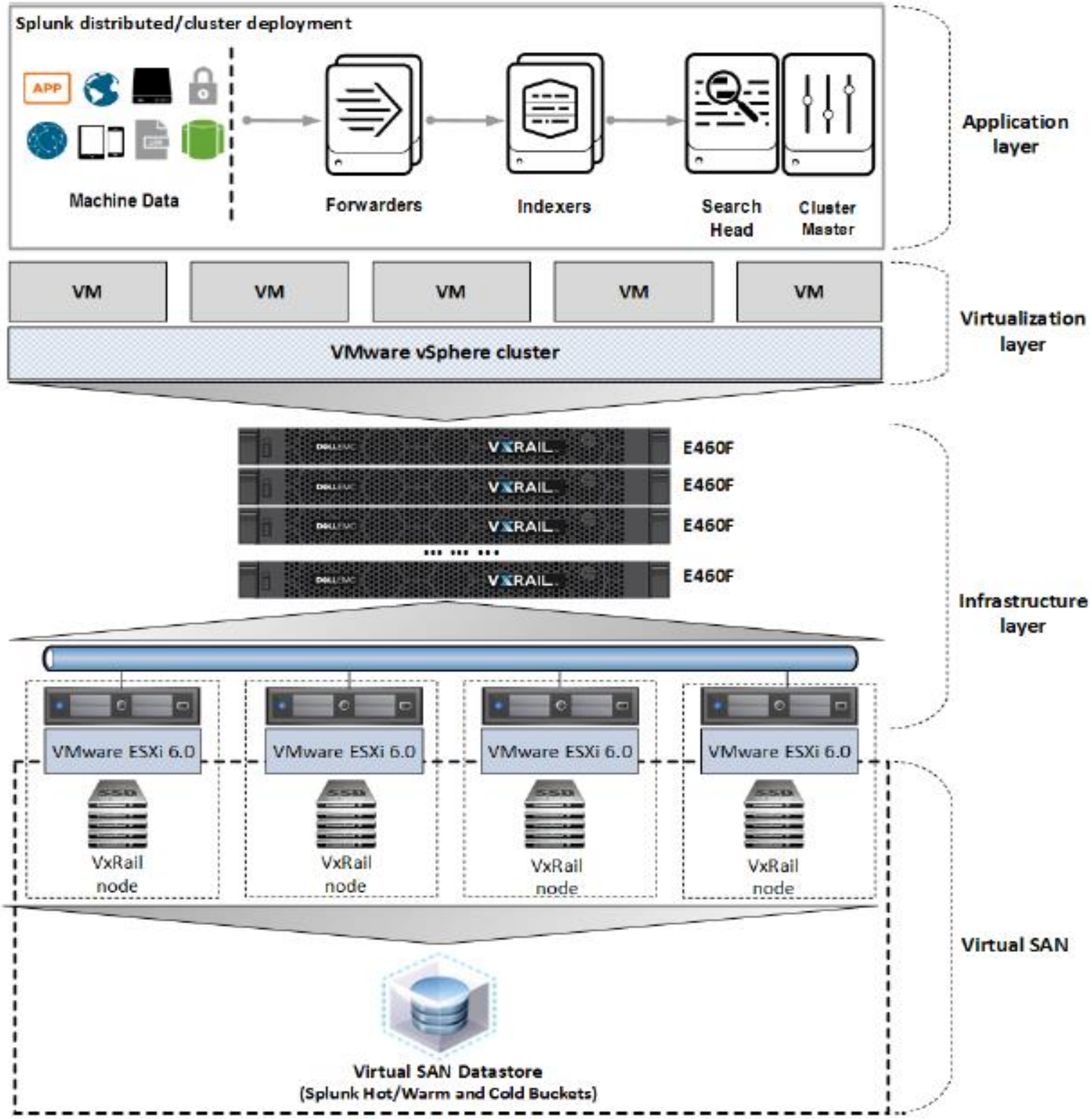


Figure 1. Splunk Enterprise on VxRail Appliance reference architecture

.

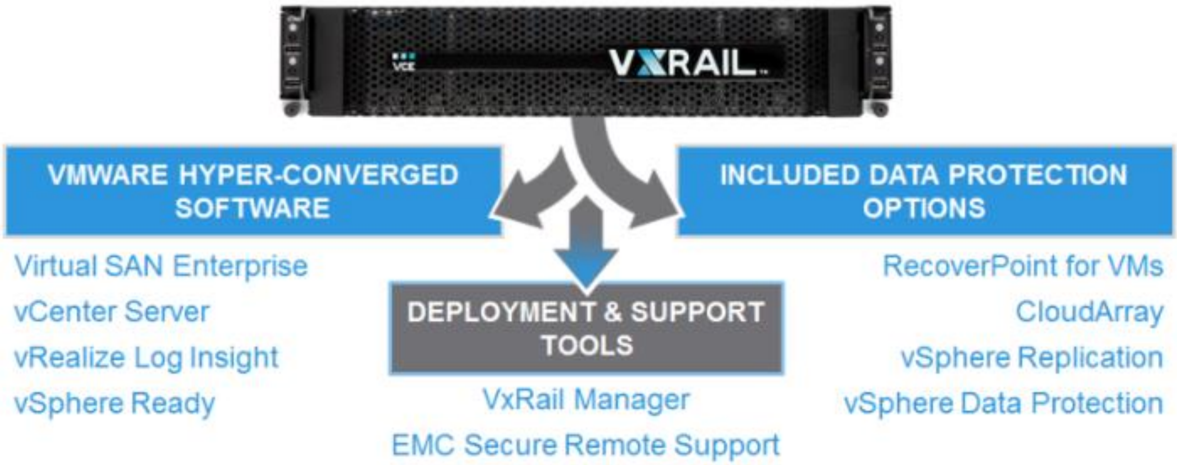
Key Components

- Virtual Storage Area Network – Virtual storage for Hot/Cold/Warm data
- Infrastructure Layer – VxRail node wthl all Flash SSD storage
- Application Layer – Dedicated VMs for Splunk

Deployment Steps

- Implement the VxRail Node cluster
- vSan Policy is prepared for Splunk disk depending on the type of data (Hot/Cold/Warm)
- Splunk indexer is deployed as per VM template
- Splunk Search head is deployed
- Splunk admin server is deployed
- Validation of Splunk deployment

VxRail Architecture Overview



Storage Component

- VxRail Appliances use VMware's vSAN software, which is fully integrated with vSphere and provides software-defined storage. vSan software is embedded at hypervisor or kernel level which make it more efficient and optimized rather than installing VSA (virtual storage appliance) on each VM and then installing vSan.

- vSan ensures the storage management policy at kernel level for seamless integration between virtual compute network and storage layers.

Compute Component

- The VxRail Appliance uses a modular, distributed system architecture based on a 1U appliance with one node that scales linearly. In addition, different options are available for compute, memory, and storage configurations to match any use cases. Choose from a range of next-generation processors, variable RAM, storage and cache capacity for flexible CPU-to-RAM-to-storage ratios.

Network Component

- VxRail is a self-contained infrastructure, not a stand-alone environment. It is intended to connect and integrate with the customer's existing data center network. The distributed cluster architecture enables independent nodes to work together as a single system. The close coupling between nodes is accomplished through IP networking connectivity.

Conclusion

With the explosion of data growth in IT data center technologies, the scope of IT challenges continues to get more disparate and complex. Big Data analytics, specifically the analysis of machine data, can help business of all sizes drive critical decisions, reduce costs, and maximize operational efficiencies to overcome these challenges.

Integration of various analytics tools with Dell Technologies solutions such as Isilon and Hyperconverged products help business leverage accurate insights from machine data that will enable optimized decision making.

References

http://www.optiodata.com/documents/optio/datasheets/dell-emc-vxrail-appliance-techbook.pdf
https://www.datacore.com/hyperconverged-infrastructure/

Dell Technologies believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

THE INFORMATION IN THIS PUBLICATION IS PROVIDED “AS IS.” DELL TECHNOLOGIES MAKES NO RESPRESENTATIONS OR WARRANTIES OF ANY KIND WITH RESPECT TO THE INFORMATION IN THIS PUBLICATION, AND SPECIFICALLY DISCLAIMS IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Use, copying and distribution of any Dell Technologies software described in this publication requires an applicable software license.