DELL Technologies

Solution Overview
Validated Design

Dell Open Data Lakehouse

Unleashing the power of AI begins with a Robust Data Foundation

Today's Situation

Capitalizing on AI begins with Data

Organizations are adopting a forward-thinking strategy, treating AI and Data as a product, to lead in innovation and value creation. In the digital age, data is a strategic asset, and AI is transformative. Maximizing AI's potential begins with a robust data foundation. It enables AI to thrive, ensuring ongoing model improvement. Investing in data ensures immediate effectiveness and long-term innovation.

The Data Challenge

In the rapidly evolving landscape of data management, organizations are grappling with a multitude of challenges including:

- People and Skills: Data teams spend only 22% of their time innovating. Traditional processes, delays, and skills shortages hinder progress.
- Data Complexities: A recurring challenge is the significant time invested by data scientists and engineers in obtaining and prepping data for AI workflows.
- Expansion of data: The proliferation of data, including multiple copies, across growing multicloud environments is hampering organizations' ability to fully leverage AI for business outcomes.
- Infrastructure: Managing data involves complex tasks, from pipelines to security compliance. Agility in infrastructure management is crucial amidst a complex tool landscape.

Why a Data Lakehouse, Why Now

For many organizations, addressing the challenges sited above involves a transformation to create a cohesive adaptable data ecosystem.

Traditional Data Stack: Today's data landscape has evolved to blend traditional data components: Data Warehouses and Data Lakes. Data Warehouses excel in structured data processing, while Data Lakes handle unstructured data with flexibility and scalability. This dual approach combines agility and governance but often leaves a gap in data agility and cost efficiency. Challenges like data integration, accessibility, and insight derivation across diverse data types can hinder decision-making and increase operational costs.

Modern Data Stack: Data Lakehouse architectures have emerged as a powerful solution that bring the best of both worlds i.e. Data Warehouses and Data Lakes. A Data Lakehouse provides:

- Unified Data Stack: Reduces data movements, simplifying the data landscape and cutting costs.
- Enhanced Data Warehouse Capabilities: Enables transaction management and time travel features, akin to traditional data warehouses.

© 2023 Dell Inc. or its subsidiaries.

Public

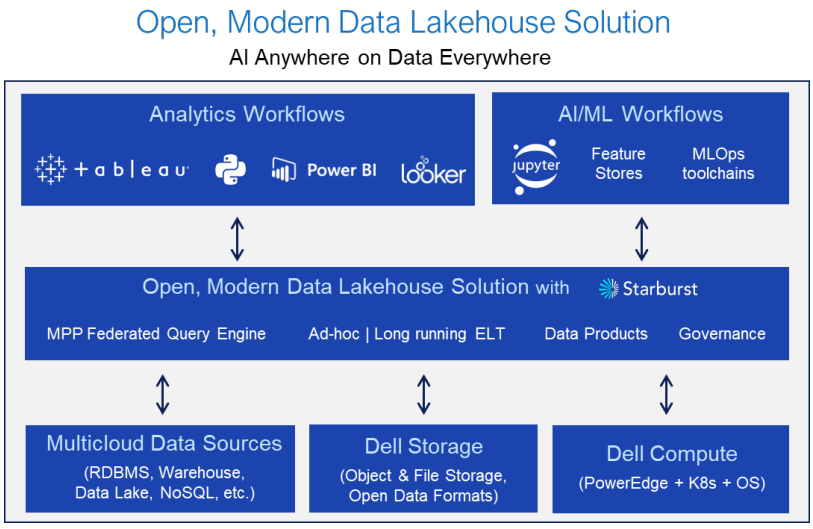
- Scalability and Performance: Separates storage and compute architecture, ensuring scalability and performance.
- Open Data Formats: Leverages open formats like Iceberg and Delta Lake, reducing vendor lock-in and offering flexibility.
- Unified Governance: Provides consistent governance and access control, ensuring data is available to the right users.
- Data Source Compatibility: Access and process data from various sources, including RDBMS, data warehouses, and data lakes, across on-premises, cloud, and edge environments.
- Data Product Creation: Facilitates the development of data products, promoting data discovery and reuse across the organization.

Dell Open Data Lakehouse

You can now partner with Dell Technologies to start your data modernization journey. At Dell, we recognize that every enterprise is at a different stage in their digital transformation journey, and we are committed to meeting you precisely where you are.

Logical View

Per the illustration below, the Dell Open Data Lakehouse sets the foundation for a modern data stack and simplifies the data landscape by minimizing data movements and consolidating data siloes.



Our customer-centric approach ensures that your IT solutions are not only tailored to your current needs but also primed for future advancements.

Key capabilities include:

- Analytics and AI/ ML Workloads: Dell partners with an open ISV Ecosystem empowering customer to plug into their existing tools to meet the varying demands across their enterprise. This includes tools for data processing, BI reporting, AIML Ops and many more.
- Open Data Lakehouse: Dell's Lakehouse engine is powered by Starburst, the leading data analytics platform that allows users to activate data wherever it resides in their enterprise.

© 2023 Dell Inc. or its subsidiaries.

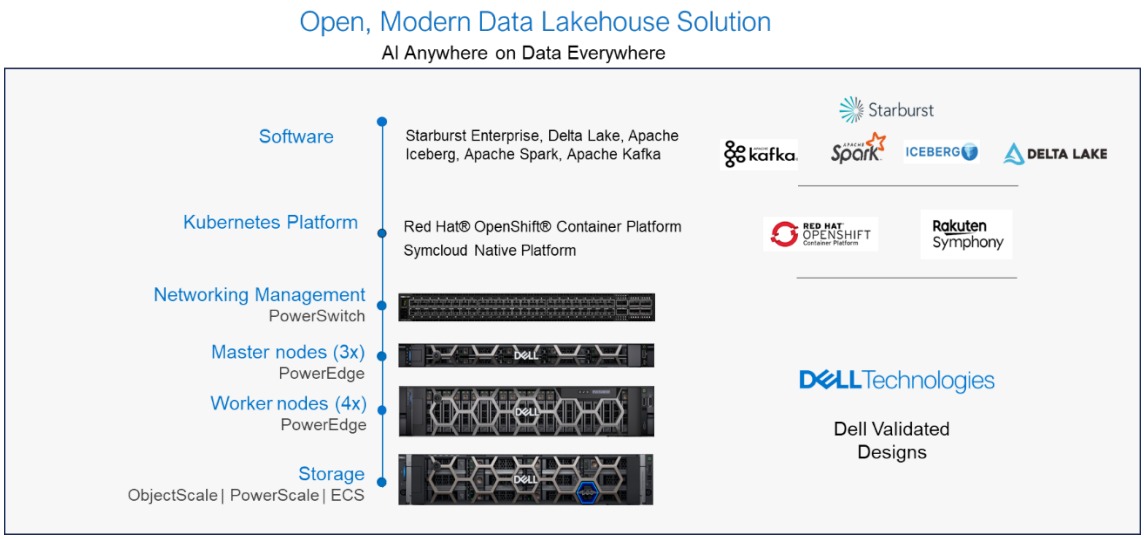
Public

The Data Lakehouse streamlines data operations by enabling in-place discovery and access to data across the enterprise, whether on-premises or in the cloud. It supports data ingestion from a variety of sources, allowing data to be materialized within the lakehouse for further processing and transformation, ensuring data is well-prepared for analysis and decision-making.

- Multicloud Data Sources: The Dell Data Lakehouse open source software and open architecture enables data federation for joining data across various sources and / or moving a data from as events, costs and needs change. It promotes data sharing through curated data product via connectors like Starburst Stargate.
- Dell Storage: Dell Object and File Storage provides a versatile data store for structured, semi-structured like CSV, XML and JSON, and unstructured data like images, .pdfs, audio, videos. Dell object storage supports read / write consistency within and across sites for handling concurrent, atomic transactions with the choice of two open table formats, Delta Lake and Iceberg. Both use parquet and Iceberg can also support Avro or ORC.
- Dell Compute: Dell PowerEdge computing and networking are finely tuned for scaling and high-performance handling of diverse data workloads. Effortless management of Kubernetes distributions like RedHat OpenShift, Rancher and Symcloud empowers seamless deployment and lifecycle management of the lakehouse stack at an enterprise level.

Dell Validated Designs

Dell Validated Designs offer substantial value by providing pre-tested, reliable solutions that streamline IT infrastructure deployment. These well-documented blueprints reduce the risk of errors, optimize system performance, and accelerate time-to-value, making them a valuable asset for efficient and successful IT projects.



Powered by Starburst together with leading open source technologies, designs are built on leading Kubernetes platforms including RedHat OpenShift container platform, SUSE Rancher and Symcloud Native Platform.

© 2023 Dell Inc. or its subsidiaries.

Public

Dell essential compute, network and storage power Dell Open Data Lakehouse validated designs include;

Component	Function
Dell Compute	The Dell PowerEdge server infrastructure delivers essential compute, memory, and storage resources vital for running customer workloads effectively. With a broad range of PowerEdge server Intel and AMD configurations available, these platforms cater to a diverse set of workloads commonly encountered in a Lakehouse architecture, ensuring versatile support for various data operations and analysis.
Dell Storage	The Lakehouse storage system has maximum flexibility leveraging Dell PowerScale with HDFS protocol or utilizing Dell ECS and Dell ObjectScale for object storage through the S3 protocol. The choice of configuration and scale is tailored to the specific workloads and the volume of data that needs support.
Dell Network	Built on the foundation of Dell PowerSwitch and SmartFabric technologies, the network is crafted to fulfill the demands of a high-performance and scalable cluster. It offers both redundancy and seamless access to comprehensive management capabilities

Why partner with Dell

Dell Technologies offers a suite of resources that bring immense value to organizations embarking on their digital transformation journeys.

Dell Executive Briefings: Tailored sessions that enable leaders to engage directly with experts and explore solutions aligning with specific business needs.

Dell Customer Solution Centers: Hands-on environments where organizations can test-drive technologies, fine-tune configurations, and validate strategies before implementation, reducing risks and enhancing success. Contact your account team to submit an engagement.

Dell Specialists for Data Management and Analytics Solutions: Dell Technologies has a dedicated team of specialists and experts who focus on data management and analytics solutions. These experts provide tailored guidance and support to organizations, helping them optimize data infrastructure, implement data governance policies, and harness data analytics for valuable insights.

Dell Global Footprint: A worldwide presence that guarantees support and scalability, facilitating seamless technology deployment and management on a global scale, making it a trusted partner for organizations worldwide.

Sizing Guidance: For more information, including sizing guidance, technical questions, or sales assistance, email analytics.assist@dell.com , or contact your Dell Technologies or authorized partner sales representative.

Learn more about Dell Data Analytics solutions

Contact a Dell Technologies Expert

Dell InfoHub resources

Join the conversation with #DellKnowsData

DELL Technologies

Copyright © 2023 Dell Inc. or its subsidiaries. All Rights Reserved. Dell and other trademarks are trademarks of Dell Inc. or its subsidiaries. Apache®, Spark®, and Kafka® are trademarks of the Apache Software Foundation or its subsidiaries in Canada, the United States, and/or other countries. Kubernetes® is a registered trademark of The Linux Foundation. Other trademarks may be the property of their respective owners. Published in the USA 11/23 Solution brief data-lakehouse-brief.

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

Appendix

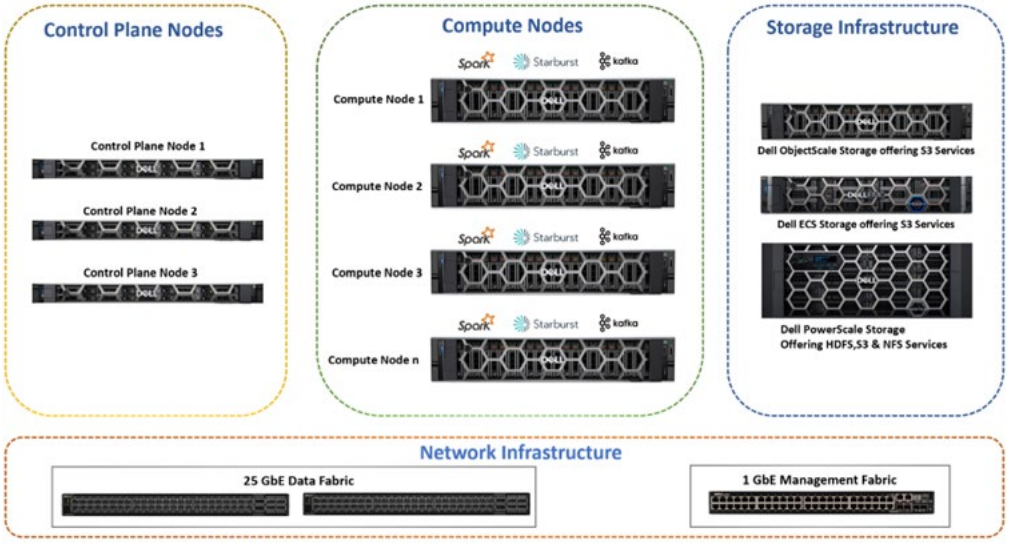
Validated Design: Dell Open Data Lakehouse on RedHat OpenShift

System Architecture & Configuration

Infrastructure Overview

Dell infrastructure provides the compute, memory, storage, and network resources to power the lakehouse platform.

- **Control nodes** : Cluster control plane enables global functions e.g. scheduling, detection and response to cluster events such as starting a new pod in a deployment.
- **Worker nodes:** Using worker nodes, Kubernetes clusters run containerized applications and provide networking services so that traffic between applications and from outside the cluster can be handled effectively.
- **Storage infrastructure:** Dell ObjectScale, PowerScale and ECS can be used as the lakehouse storage layer.
- **Network infrastructure:** Network infrastructure provides the required connectivity between the server and storage infrastructure and the on-premises network.



Server Infrastructure

The server infrastructure provides compute, memory, and some of the storage resources that are required to run customer workloads. A wide variety of PowerEdge server configurations are possible. The recommendations here support a wide variety of workloads typical in a lakehouse architecture implementation.

Lakehouse control plane node configuration

Three control plane nodes are required for production clusters to provide high availability for the control plane. Memory, storage, and processor have been sized to support all the required services in a production deployment.

Machine function	Component
Platform	PowerEdge R660 server
Chassis	2.5 in chassis with up to 10 hard drives (SAS or SATA), two CPU slots, and PERC 12

© 2023 Dell Inc. or its subsidiaries.

Public

Chassis configuration	Riser configuration 2, three 16-channel, low-profile slots (two Gen5 and one Gen4)
Power supply	Dual hot-plug, fault-tolerant (1+1), 1100 W mixed mode (100-240 Vac), Titanium, normal airflow (NAF) power supplies
Processor	Intel Xeon Gold 6426Y 2.5 G, 16 C/32 T, 16 GT/s, 38 M cache, turbo, HT (185 W) DDR5-4800
Memory capacity	128 GB (eight 16 GB RDIMM, 3200 MT/s, dual rank)
Internal RAID storage controllers	Dell PERC H965i with rear load bracket
Disk—SSD	Two 1.6 TB 2.5 in hot-plug, SAS, mixed-use, up to 24 Gbps 512e Federal Information Processing Standard (FIPS), three drive writes per day (DWPD) SSDs
Boot optimized storage cards	BOSS-N1 controller card + with two M.2 960 GB SSDs (RAID 1)
Network interface controllers	NVIDIA ConnectX-6 Lx dual port 10/25 GbE SFP28 adapter, PCIe low profile

Lakehouse worker node configuration

Lakehouse worker nodes support the platform runtime services and customer workloads.

Machine function	Component
Platform	PowerEdge R660 server
Chassis	2.5 in chassis with up to eight SAS or SATA hard drives, three PCIe slots, and two CPUs
Chassis configuration	Riser configuration 2, three low-profile 16-channel slots, (two Gen5 and one (Gen4)
Power supply	Dual hot-plug, fully redundant (1+1), 2400 W, mixed mode power supplies
Processor	Intel® Xeon® Gold 6448Y 2.1G, 32C/64T, 16GT/s, 60M Cache, Turbo, HT (225W) DDR5-4800
Memory capacity	768 GB (Twelve 64 GB RDIMM, 4800 MT/s, dual rank)
Internal RAID storage controllers	Dell PERC HBA355i with rear load bracket
Disk—SSD	Two 1.6TB SSD SAS Mixed Use up to 24Gbps 512e 2.5in Hot-Plug
Boot optimized storage cards	BOSS-N1 controller card + with two M.2 960 GB NVMEs (RAID 1)
Network interface controllers	NVIDIA ConnectX-6 Lx dual port 10/25 GbE SFP28 adapter, PCIe low profile

General-purpose lakehouse worker node volumes

Usage	Volume type	Physical disks	Volume ID
Operating system	RAID 1	Two M.2 960 GB NVMEs	0

© 2023 Dell Inc. or its subsidiaries.

Public

Storage Infrastructure

The Lakehouse storage system can use PowerScale with HDFS protocol, or ECS or Dell ObjectScale for object storage with S3 protocol. The configuration and scale would differ for every customer depending on the nature of workloads, volume to be supported etc.

PowerScale H7000

Machine function	Component
Model	PowerScale H7000 (hybrid)
Chassis	4U node
Nodes per chassis	Four
Node storage	Twenty 12 TB 3.5 in. 4 kn SATA hard drives
Node cache	Two 3.2 TB SSDs
Usable capacity per chassis	600 TB
Front-end networking	Two 25 GbE (SFP28)
Infrastructure (back-end) networking	Two InfiniBand QDR or two 40 GbE (QSFP+)
Operating system	OneFS 9.5.0.2

ECS EX500

The ECS EX500 configuration provides a good balance of storage density and performance for lakehouse usage.

Machine function	Component
Model	ECS EX500
Chassis	2U node
Nodes per rack	16
Node storage	960 GB SSD
Node cache	N/A
Usable capacity per chassis	Slightly less than 384 TB
Front-end networking	Two 25 GbE (SFP28)
Infrastructure (back-end) networking	Two 25 GbE (SFP28)

© 2023 Dell Inc. or its subsidiaries.

Public

ECS EXF900

The ECS EXF900 configuration is an all-flash configuration and provides the highest performance for lakehouse usage.

Machine function	Component
Model	ECS EXF900
Chassis	2U node
Nodes per rack	16
Node storage	184 TB (twenty-four 7.68 TB NVMe drives)
Node cache	N/A
Usable capacity per chassis	Slightly less than 184 TB
Front-end networking	Two 25 GbE (SFP28)
Infrastructure (back-end) networking	Two 25 GbE (SFP28)

Dell ObjectScale

Dell ObjectScale is a software-defined solution for S3-compatible enterprise grade object storage. ObjectScale uses a containerized architecture to deliver enterprise-class, high-performance object storage in a Kubernetes-native package.

Machine function	Component
Model	Dell ObjectScale All Flash
Platform	Dell PowerEdge R760 Server
Nodes per rack	16
Chassis	2.5" Chassis with up to 24 NVMe Direct Drives, 2 CPU
Chassis configuration	Riser Config 3, Half Length, 2x8 FH Slots (Gen4), 2x16 FH Slots (Gen5), 2x16 LP Slots (Gen4)
Power supply	Dual, Hot-plug Fully Redundant Power Supply (1+1) 1100W
Processor	Intel® Xeon® Gold 6426Y 2.5G, 16C/32T, 16GT/s, 38M Cache, Turbo, HT (185W) DDR5-4800
Memory capacity	512 GB (sixteen 32GB RDIMM, 4800MT/s Dual Rank)
Internal RAID storage controllers	C30, No RAID for NVME chassis
Disk—SSD	Twenty four 6.4TB Enterprise NVMe Mixed Use AG Drive U.2 Gen4 with carrier
Boot optimized storage cards	BOSS-N1 controller card + with two M.2 960 GB SSDs (RAID 1)
Network interface controllers	NVIDIA ConnectX-6 Lx dual port 10/25 GbE SFP28 adapter, PCIe low profile
Node storage	153.6 TB (twenty-four 6.84 TB NVMe drives)
Front-end networking	Two 25 GbE (SFP28)

© 2023 Dell Inc. or its subsidiaries.

Public

Network Components

The network is designed to meet the needs of a high performance and scalable cluster, while providing redundancy and access to management capabilities. The architecture is a leaf and spine model that is based on Ethernet networking technologies.

- It uses PowerSwitch S5248F-ON switches for the leaves and PowerSwitch Z9432F-ON switches for the spine.
- The management network uses a PowerSwitch S3148 1 GbE switch for iDRAC connectivity and chassis management.
- The switches run Dell SmartFabric OS10. SmartFabric OS10 enables multilayered disaggregation of network functions that are layered on an open-source Linux-based operating system.
- The VLT configuration in this design uses four 100 GbE ports between each Top of Rack (ToR) switch. The remaining 100 GbE ports can be used for high-speed connectivity to spine switches, or directly to the data center core network infrastructure.

Container Platform

The lakehouse platform can be built using any of the existing Kubernetes distributions. The validation in this paper was done using RedHat OpenShift Container Platform 4.12.

Software Components

- Apache Spark 3.4.1
- Hadoop client libraries 3.3.6 (for HDFS access)
- Hadoop AWS libraries 3.3.6 (for S3 access)
- Delta Lake 2.4.0 & 3.0.0rc1 libraries
- Iceberg 1.3.1

For additional technical information:

- Dell Open Data Lakehouse with RedHat Open Shift Platform: White Paper
- Dell Open Data Lakehouse with RedHat OpenShift Platform: Design Guide

© 2023 Dell Inc. or its subsidiaries.

Public