DELL Technologies

Spec Sheet

Dell Data Lakehouse

Author: Kirankumar Bhusanurmath, DA / AI Specialist | TME, Dell Technologies

Dell Data Lakehouse Overview

The Dell Data Lakehouse provides the best experience for a modern data platform. A fully integrated data lakehouse, built on Dell hardware with a full-service software suite. Its distributed query processing enables organizations to federate data with minimal data movement. Additionally, they can centralize their data estate into a modern data lake and benefit from performant SQL access to this data. The Dell Data Lakehouse employs the Dell Data Analytics Engine powered by Starburst, a high performance massively parallel query engine. It enables the discovery, querying, and processing of all enterprise data, irrespective of location and data sources.

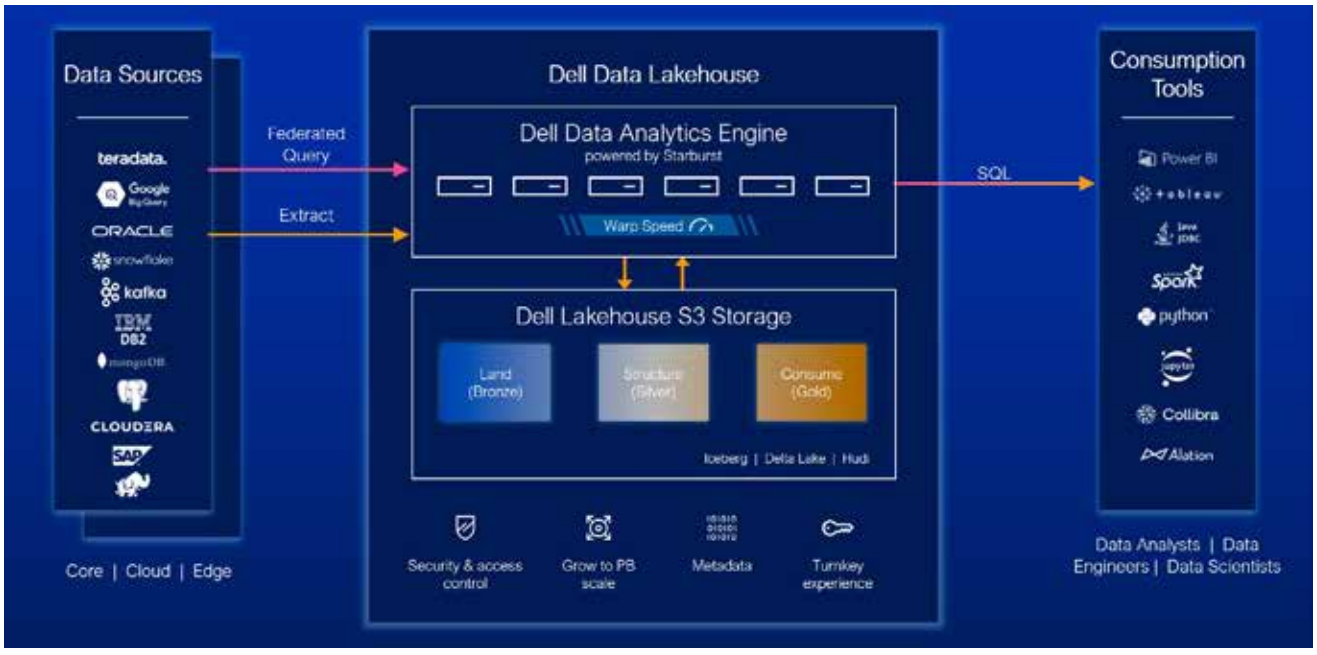

Figure 1. Dell Data Lakehouse Diagram

Dell Data Lakehouse Components

The Dell Data Lakehouse solution is made up of four key components.

Hardware Components

1. Compute Nodes aka Dell Data Analytics Engine (DDAE660)
2. Storage Nodes Dell ECS or ObjectScale or PowerScale storage cluster
3. Dell Networking equipment S5248F-ON and Z9432F -ON

SW Components

1. Dell Data Analytics Engine, powered by Starburst
2. Dell Data Lakehouse System Software

Hardware Components

Compute Nodes aka DDAE660

The DDAE 660 compute nodes are built on the powerful and reliable Dell PowerEdge R660 server (1U). The cluster of these nodes is connected to connected to network switches, such as Dell PowerSwitch S5248F-ON and Dell PowerSwitch Z9432F-ON, or similar customer-provided network switches.

Storage Nodes

Dell ECS

Dell ECS, the world's most cybersecure object storage1, gives you unmatched scalability, performance, resilience, and economics. ECS delivers rich S3-compatibility on a globally distributed architecture, empowering organizations to support enterprise workloads such as AI (Artificial Intelligence), analytics, and archiving at scale. ECS customers are reducing TCO (Total Cost of Ownership) up to 76% over public cloud.

Dell ObjectScale

ObjectScale is high-performance containerized object storage built for the toughest applications and workloads—Generative AI, analytics, and more. Innovate faster, at any scale, with a global namespace, strong S3 compatibility, and enterprise-class security that is ready on day one. Expanding its software-defined options, ObjectScale is also available as the world's most powerful object storage appliance purpose-built for Kubernetes.

Dell PowerScale

The world's most flexible, secure, and efficient scale-out file storage

Network switches

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.

Sowftware Components

Dell Data Analytics Engine powered by Starburst.

Dell Data Analytics Engine contains an analytics query engine powered by Starburst. Dell Data Analytics Engine is a fully supported and enterprise-grade distributed SQL query engine designed for high-performance analytics. It allows users to query large amounts of data stored in various data sources throughout an organization using standard SQL syntax. One of the key features of Dell Data Analytics Engine is its ability to run queries across different data sources simultaneously, in the same query. These sources include relational databases, NoSQL databases, object storage systems, and more.

Response times are fast enough to support real-time analysis. The Dell Data Analytics Engine can also execute high performance queries on raw data stored in a data lake in file formats such as Parquet and ORC. Users can use this engine to materialize data in an object store in open table formats (e.g., Iceberg, Delta Lake) to form an open, modern data lakehouse.

With the integrated query engine, administrators can implement a layer on top of data that abstracts away details on location, connectivity, language variations, and API. This layer of abstraction is critical to simplify data analytics over a diverse set of data sources. This engine comes with access control built-in and other advanced security features.

Dell Data Lakehouse System Software

The Dell Data Lakehouse System software is the central nervous system of the Dell Data Lakehouse. It simplifies lifecycle management of the entire stack, drives down IT OpEx with prebuilt automation, provides visibility into the cluster health, enables easy upgrades and patches, lets admin control all aspects of the cluster from one convenient control center.

The Dell Data Lakehouse offers three user interfaces tailored to different users' needs: one for lakehouse cluster management, another for user management, and a third for data querying and analytical tasks via the Dell Data Analytics Engine.

Dell Data Lakehouse Components Specifications

Dell Data Analytics Engine Specifications

Table 1. Dell Data Analytics Engine Specifications

Item	Dell Data Analytics Engine Specification	Description
Framework	Distributed	In distributed computing, multiple devices or systems handle processing instead of relying on a single central device. Each device or system possesses its processing capabilities and may store and manage its data. They collaborate to perform tasks and share resources, without any single device acting as the central hub.
Connectivity to Object Storage	Yes	DDAE effortlessly connects to S3 compliant object storage platforms, enabling streamlined data querying and analytics directly from Dell Object Storage
Connectivity to Operational RDBMS	Yes	DDAE facilitates seamless connectivity to operational RDBMS (Relational Database Management Systems), allowing efficient querying and analysis directly from relational databases such as MySQL, PostgreSQL, Oracle, and SQL Server for example.
Connectivity to Nonrelational DBs	Yes	DDAE enables seamless connectivity to non-relational databases (NoSQL databases), empowering efficient querying and analysis directly from data stores such as MongoDB, Cassandra, Elasticsearch, and Apache HBase for example.
Connectivity to Streaming Sources	Yes	DDAE provides robust connectivity to streaming sources, enabling real-time data querying and analysis from platforms such as Apache Kafka and Amazon Kinesis.
Connectivity to Public Cloud Sources	Yes	DDAE offers seamless connectivity to public cloud sources, facilitating efficient querying and analysis directly from platforms like Amazon Web Services (AWS), Google Cloud Platform (GCP), Microsoft Azure, and other cloud services
Data Federation	Yes	Data federation in DDAE involves integrating and querying data from multiple disparate sources without the need for data movement or duplication. As a distributed query engine, DDAE enables unified access to data across various databases, file systems, and cloud storage platforms, streamlining analytics workflows and providing a cohesive view of the data landscape for analysis and decision-making purposes.

Connectors	50+	Connectors in DDAE serve as interfaces to various data sources, enabling seamless interaction and query execution across relational databases, non-relational databases, streaming platforms, and cloud storage services. These connectors are crucial components that allow DDAE to access and process data from diverse sources efficiently, empowering organizations to perform comprehensive analytics and gain insights from their data ecosystem
Concurrency	Yes	Concurrency in DDAE refers to its capability to handle multiple queries simultaneously, efficiently managing resources and ensuring optimal performance even during peak workloads. By leveraging its distributed architecture and query optimization techniques, DDAE effectively handles concurrent queries from multiple users or applications, maximizing throughput and minimizing query latency for enhanced analytics productivity.
Dependency on Hadoop	No	DDAE does not have a direct dependency on Hadoop. While DDAE's query engine, Trino, originated as a project within the Hadoop ecosystem, it has evolved into a standalone distributed SQL query engine. Trino can run on various platforms independently of Hadoop, including cloud environments, Kubernetes, and traditional on-premises setups. It provides connectors to various data sources, allowing users to query data stored in Hadoop Distributed File System (HDFS) or Hadoop-compatible file systems like Amazon S3 or Azure Data Lake Storage alongside other data sources without requiring a full Hadoop deployment. Therefore, while Trino can integrate with Hadoop ecosystems if needed, it does not rely on Hadoop to function.
Requirement for YARN	No	DDAE does not require Apache YARN (Yet Another Resource Negotiator) for resource management. Unlike some other tools in the Hadoop ecosystem, DDAE operates independently and manages its resources internally. It utilizes its own resource management and scheduling mechanisms to execute queries efficiently across its distributed architecture without relying on YARN. Therefore, deploying DDAE does not necessitate the presence of YARN in the environment.
Ranger integration	Yes	DDAE can integrate with Apache Ranger for centralized authorization and access control management. By configuring Ranger policies, administrators can define fine-grained access rules for Trino queries, ensuring data security and compliance. Ranger provides a centralized interface for managing permissions across various data sources, including Hadoop components, cloud storage platforms, and databases, thus enabling consistent security policies enforcement for Trino queries across the entire data ecosystem.
Analytics Workloads	Yes	Analytics workloads in DDAE involve processing large volumes of data to extract insights and make data-driven decisions. DDAE's distributed query engine excels at executing complex analytical queries across diverse data sources efficiently. Whether performing ad-hoc queries, interactive data exploration, or batch analytics, DDAE enables organizations to analyze structured and unstructured data from relational databases, cloud storage, streaming platforms, and more, empowering users to derive valuable insights for strategic decision-making and business optimization.
Open Table Format support	Yes - Delta Lake, Hudi, Iceberg	DDAE supports querying data stored in the Open Table Format, enabling seamless integration and analysis of tabular data structured according to this format. With its versatile connector architecture, DDAE can access and process Open Table Formatted data alongside other data sources, offering users the flexibility to perform comprehensive analytics across diverse datasets.
Cross-cloud Data Access	Yes	Cross-cloud data access in DDAE facilitates querying and analyzing data stored across multiple cloud platforms, such as Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure. DDAE's connectors for various cloud storage services enable unified access to data regardless of its location, empowering organizations to leverage data from different cloud providers seamlessly for comprehensive analytics and insights generation. This capability enhances flexibility, scalability, and agility in data analytics workflows across heterogeneous cloud environments.
Cross-cloud analytics	Yes	DDAE provides connectivity to other DDAE cluster through stargate connector enabling cross-cloud analytics.

Language Support	SQL, Python	DDAE provides extensive language support, allowing users to interact with the query engine using various programming languages and query dialects. Its SQL-compliant interface enables users to write standard SQL queries for data retrieval and manipulation. Additionally, DDAE supports other languages such as Python, this enhances DDAE's versatility and interoperability, catering to diverse user preferences and requirements in data analytics and application development.
Data Catalog	Yes	The data catalog in DDAE serves as a centralized repository for metadata about the available data sources, tables, columns, and other relevant information. It provides a comprehensive view of the data landscape accessible to DDAE, enabling users to discover, understand, and query data efficiently. DDAE's data catalog supports various metadata storage systems, including Hive Metastore, AWS Glue Catalog, and PostgreSQL, allowing organizations to manage metadata according to their preferences and requirements. This centralized metadata management facilitates data governance, data lineage tracking, and collaboration among data consumers and analysts.
Policies	Yes	Policies in DDAE, often managed through tools like Apache Ranger, define rules and permissions for accessing and manipulating data within the system. These policies enable administrators to enforce fine-grained access control, specifying which users or groups have permission to perform specific actions on datasets or tables. By implementing policies, organizations can ensure data security, compliance with regulations, and adherence to internal data governance standards. This granular control over data access helps mitigate risks and safeguard sensitive information while facilitating authorized data usage for analytics and decision-making purposes.
Access Control	Yes	Access control in DDAE regulates users' and applications' permissions to interact with data and perform actions within the system. Through mechanisms like role-based access control (RBAC) and permissions management, administrators can define who can access which data sources, execute queries, create, or modify objects, and perform administrative tasks. By enforcing access control policies, organizations can protect sensitive data, ensure compliance with regulatory requirements, and maintain data governance standards. Access control mechanisms in DDAE provide granular control over data access, promoting security, integrity, and confidentiality in data processing and analysis workflows.
Data Sharing	Yes	Data sharing in DDAE involves facilitating collaboration and data exchange among users, teams, or organizations by enabling seamless access to shared datasets. DDAE provides Data products which are served by a domain and consumed by downstream users to produce business value. DDAE's federated query capabilities allow users to query and analyze data from multiple sources, including shared datasets hosted on different platforms or environments and create sharable Data products. Additionally, DDAE supports features like views, federated queries, and connectors to enable data sharing across various data repositories and systems. This promotes agility, flexibility, and efficiency in data collaboration efforts, empowering organizations to leverage shared data assets for analytics, reporting, and decision-making purposes
Security Authentication Types	Password Files, LDAP, LDAP Group Provider, OAuth2, OAuth2 over HTTPS	DDAE offers security authentication via OAuth2 or traditional username/password authentication, optionally backed by LDAP as a user repository. These mechanisms ensure secure access to the system and data, facilitating integration with various identity providers and centralized user credential management. Organizations can implement authentication methods according to their security standards, ensuring robust data protection within the Dell Data Lakehouse environment.
User Interface	CLI, Query Editor	DDAE provides a user-friendly interface for executing SQL queries and managing data analytics workflows. Its web-based UI, commonly accessed through a web browser, offers an intuitive query editor with syntax highlighting and autocomplete features. Users can monitor query execution progress, view query results, and access system information through interactive dashboards. Additionally, DDAE supports command-line interfaces (CLI) for advanced users, enabling efficient interaction with the system through terminal-based commands. This comprehensive UI enhances usability and productivity, empowering users to perform complex data analysis tasks with ease

REST API	Yes	DDAE exposes a comprehensive REST API that allows programmatic interaction with the query engine. Through this API, users can submit SQL queries, monitor query execution, retrieve query results, and manage system configurations programmatically. The REST API enables seamless integration of DDAE with external applications, automation of query workflows, and development of custom solutions for managing and interacting with Lakehouse clusters. This RESTful interface enhances flexibility and extensibility, empowering users to build scalable and efficient data analytics pipelines and applications leveraging the capabilities of DDAE.
Support for Materialized Views	Yes	DDAE offers support for materialized views, providing a mechanism to precompute and store the results of complex or frequently used queries. By creating materialized views, users can improve query performance and reduce latency by accessing precomputed data instead of re-executing the original query. DDAE's materialized views feature allows users to define and maintain materialized views using SQL syntax, enabling efficient data summarization, aggregation, and denormalization for enhanced query performance and analytical capabilities. This functionality enhances DADE's versatility and scalability, making it suitable for a wide range of data analytics use cases.
Deployment	Dell Data Lakehouse only	Led by Dell ProDeploy Teams
Integration with other tools	Yes	DDAE seamlessly integrates with a diverse range of tools and frameworks, facilitating comprehensive data analytics solutions. It supports integration with popular BI tools, data science platforms, ETL (Extract, Transform, Load) tools, data warehouses, streaming platforms, and storage systems. Additionally, DDAE integrates with security and governance tools, ensuring fine-grained access control and auditing capabilities for enhanced data security and compliance
Warp Speed	Yes	Dell Data Analytics Engine's Warp Speed transparently adds an indexing and caching layer to enable higher performance. You can take advantage of the performance improvements by enabling Warp Speed. Your DDAE cluster nodes are provisioned with suitable hardware and configurations to setup Warp Speed utility connector for any catalog accessing object storage with the Hive, Iceberg, or Delta Lake connector.
Third-party Object Storage	Yes	DDAE allows customer to use other S3-protocol compatible storage devices with DDAE. The list of storage options is constantly evolving. To get the latest information, please contact your Dell sales representative to determine if your S3-compatible storage is supported with DDAE or not.
Internal Encryption	Yes	DDAE provides end-to-end encryption for internal DDAE components, including Coordinator, Worker, Cache Service, internal Hive meta store and database pods on the customer LAN. As the communication includes customer query data and data from the data sources, encryption is an important implementation detail. This is an optional feature and can be toggled on the cluster configuration page in the Lakehouse Admin UI. Remember – internal encryption can have a performance impact on the overall performance of the cluster.
Kerberos on External Hive Metastore	Yes	DDAE Catalog supports Kerberos on external Hive Metastore of federated data sources.
DDAE HMS external Query Engine access	Yes	External engines such as Spark and Flink can securely access metadata in the Dell Data Lakehouse to enable data discovery, processing and governance. Admins can also choose to enable this access with and without TLS and Kerberos depending on the level of security required.
Support custom Trino connector	Yes	Customers can bring their own custom Trino connector. From databases like Cassandra, MariaDB and Redis, to other sources such as Google Sheets and local files or even a proprietary application within the customer environment, users can now expand their access further into their distributed data silos.

Dell Data Lakehouse System Software Specifications

Table 2. Dell Data Lakehouse System Software Specifications

Item	Dell Data Lakehouse	Description
Security Authentication Types	Keycloak authentication, LDAP	Dell Data Lakehouse System Software includes built-in oAuth2/OIDC authentication, as well as basic authentication on select interfaces without any limits on number of users supported.
User Interface	Admin console	Dell Data Lakehouse System Software functionality can be accessed via a web browser-based UI to monitor and manage the Dell Data Lakehouse cluster.
Dial Home	Support Assist	Dell Data Lakehouse System Software allows the Dell Data Lakehouse to securely connect to Dell Services systems to report issues and proactively create service tickets. Additionally, it allows Dell services to remotely connect to the appliance over SSH to collect logs and troubleshoot issues to shorten overall ticket resolution time.
Monitoring	HW/SW Alerts	Users can use the browser-based interface to view node configuration, usage trends and detailed software component version information. The Dell Data Lakehouse System Software comes with pre-integrated alerts out of the box which can be viewed easily from the UI.
Logging	Yes	Support Materials or Data Collects (DCs) are the most common logs for analysis and troubleshooting. Log collection is an important functionality provided by Dell Data Lakehouse System Software that can bring great insight to Dell ProSupport, Dell Engineering and Development teams of diagnose the components inside Dell Data Lakehouse and help in troubleshooting any kind of technical issues. This functionality is available to remote service engineers through Support Connectivity or Support Assist that can be enabled at the time of install. It enables Dell's remote support teams to dial in and check System health and collect necessary logs for further troubleshooting.
Backup/Recovery	Yes	The Dell Data Lakehouse System Software allows for backup and restore of the cluster. This can be critical during upgrades to ensure there is no configuration or data loss. The key components are therefore the ETCD database (where Kubernetes objects/resources reside) and the internal database (where user authentication data from Keycloak and Hive Metastore resides). A backup can be generated either periodically (enabled at the time of install) or at any given point in time backup via a REST API. To generate a backup, an NFS export or CIFS external share is required. The restore procedure then restores this backup onto the cluster.
High Availability	Yes	Dell Data Lakehouse supports High Availability at multiple levels. At the hardware level, the compute nodes within the cluster provide High Availability across Hardware parts. Hardware redundancy for the product is inherited from PowerEdge server platform which are known for their highly available hardware and configurations. For example, hard drives within the compute clusters are configured to be fault tolerant. Other components like Power Supply, Memory, Network Interface Cards, cooling fans and network switches are all built to be fault tolerant. At the software level, the Dell Data Lakehouse is designed to handle node failures. For example, there are 3 control plane nodes in every cluster designed for high availability. If a control plane node fails, the cluster can continue to function until the control plane node is replaced. Similarly, if a coordinator node fails, the coordinator service can be transferred over to a worker node temporarily to maintain system functionality until the coordinator node is replaced. Finally, if a worker node fails, the coordinator node can redistribute the queries to other worker nodes until the worker node is replaced. Within the Dell Data Analytics Engine, the Fault Tolerant Execution mode can make the above scenario of a worker node failure seamless to the end user. To enable this, intermediate stages of query processing must be stored in an object store such as Dell ECS. For more details, please reach out to your account representative.

License Management	Yes	The Dell Data Lakehouse System Software UI provides easy access to license details and cluster IDs that are useful for support ticket tracking. The system also sends proactive alerts to indicate licenses are close to expiration so teams can renew in time and avoid any disruption.
Connector Management	Yes	The Dell Data Lakehouse System Software includes a wizard that helps admins add or manage external data source connections and configurations for use in Dell Data Analytics Engine. This replaces the command line interface commonly used today. Admins can also upload additional drivers to customize or enhance the functionality of connectors.
Cluster Management	Yes	Dell Data Lakehouse System Software includes pre-built integrations with underlying infrastructure for cluster management activities such as resource configuration, cluster health monitoring, alerts, enabling/disabling fault tolerant execution, managing data products, assigning/unassigning data sources, etc.
ECS Object Storage Management	Yes	With the Dell Data Lakehouse System Software, users can easily configure S3 storage connection which will be used for fault tolerant execution, backup and restore as well as storing materialized views.
User management	Yes	Dell Data Lakehouse System Software is pre-integrated with Keycloak, an open-source Identity and Access Management solution, to enable authentication and authorization between various internal components and user facing components such as the user interface, REST API and Dell Data Analytics Engine endpoints. Keycloak can also be integrated with external identity providers.
TLS Encryption	Yes	Communication between external data sources or clients and the Dell Data Lakehouse is encrypted with TLS 1.2. Communication internally within the Dell Data Lakehouse components is currently not encrypted.
Alerts	Yes	Dell Data Lakehouse comes pre-built with an extensive list of alerts.
Health Checkup	Yes	DDAE The Health Checkup provides a mechanism to establish the healthiness state of the cluster considering the following components/end points of the cluster: 1. Node Hardware/Firmware 2. OS installation (version and readiness) 3. Kubernetes installation (version and readiness) 4. Dell Data Analytics Engine (readiness) 5. Underlying Management Software 6. Connectivity (network and GW) 7. Connectivity (DNS) 8. Connectivity (ECS or other Object Store) 9. Database readiness 10. Licenses (validity, at runtime) 11. Remote connectivity (ESE) 12. NTP Configuration

License Renewal	Yes	Dell Data Lakehouse system software lets customers to upload and update the licenses for Lakehouse system software and DDAE in the UI.
Toggle Remote ESE/ Telemetry	Yes	Dell Data Lakehouse can collect and send critical Hardware system failure alerts to Dell support to provide information on failure states or pending failure conditions. This feature can be enabled or disabled in the Dell Data Lakehouse system software.
Telemetry Node-Down event alerting	Yes	Dell Data Lakehouse System Software is capable of identifying and sending node-down failure alerts to Dell support and the local admin UI.
Self-service Certificate Updates	Yes	Admins can update CA certificates, TLS certificates and TLS keys directly in the admin control plane.

Compute Nodes aka DDAE660

There are two types of nodes in a Dell Data Lakehouse compute cluster:

- Control Plane Nodes run Lakehouse System Software.
- Worker and Coordinator Nodes run the Dell Data Analytics Engine

Table 3. Compute Nodes aka DDAE660

Item	Dell Data Lakehouse	Description
Server Model	Dell Data Analytics Engine Compute 660 based on PowerEdge Rack Server	Dell Data Analytics Engine Compute 660 based on PowerEdge Rack Server
CPU (dual socket)	Intel Xeon Gold 5416S 2G, 16C/32T, 16GT/s, 30M Cache, Turbo, HT (150W) DDR5-4400	Intel Xeon Gold 5416S 2G, 16C/32T, 16GT/s, 30M Cache, Turbo, HT (150W) DDR5-4400
Total CPU Cores (Threads)	32 (64 Hyper threads)	32 (64 Hyper threads)
Memory	128 GB at 4800 MT/s	256 GB at 4800 MT/s
Hard Drives	960GB SSD (RAID 5)	480GB SSD SATA (RAID 1) 2 X 3.84TB SSD SAS Read Intensive up to 24Gbps SED FIPS-140 512e 2.5in Hot-Plug AG Drive (For Warp Speed)
Network OCP (SFP)	Intel E810-XXV Dual Port 10/25GbE SFP28, OCP NIC 3.0	Intel E810-XXV Dual Port 10/25GbE SFP28, OCP NIC 3.0
Network PCIe (SFP)	Intel E810-XXV Dual Port 10/25GbE SFP28 Adapter, PCIe Low Profile	Intel E810-XXV Dual Port 10/25GbE SFP28 Adapter, PCIe Low Profile

Storage Nodes

Primary storage to the Dell Data Lakehouse can be either Dell ECS or ObjectScale storage or Dell PowerScale storage.

Dell ObjectScale, ECS, and PowerScale are deployed as cluster-level systems. The node recommendations here can be used as guidance for new clusters, verification of compatibility with existing clusters, or expansion of existing clusters.

ECS Node

Dell Technologies recommends the configurations in ECS EX500 node configuration or ECS EXF900 node configuration for storage in clusters using ECS for their primary lakehouse storage using the s3a:// protocol.

Table 4. ECS EX500 and EXF900 node configuration

Model	ECS EX500	ECS EXF900
Model ECS	EX500	EXF900
Chassis	2U node	2U node
Nodes per rack	16	16
Node storage	384 TB (twenty-four 16 TB NLSAS drives)	184 TB (twenty-four 7.68 TB NVMe drives)
Node cache	960 GB SSD	N/A
Usable capacity per chassis	Slightly less than 384 TB	Slightly less than 184 TB
Front-end networking	Two 25 GbE (SFP28)	Two 25 GbE (SFP28)
Infrastructure (back-end) networking	Two 25 GbE (SFP28)	Two 25 GbE (SFP28)

The ECS EX500 configuration provides a good balance of storage density and performance for lakehouse usage. And the ECS EXF900 configuration is an all-flash configuration and provides the highest performance for lakehouse usage.

ObjectScale infrastructure

Dell Technologies recommends the configuration in ObjectScale all flash configuration for storage clusters that use ObjectScale for primary lakehouse storage using the s3a:// protocol.

Customers can use either of the following options to deploy ObjectScale:

- Software-defined storage to deploy ObjectScale software on Red Hat OpenShift.
- An XF960 appliance that is based on the latest generation of Dell PowerEdge servers.

Table 5. ObjectScale XF960 all flash configuration

Machine Function	Component
Platform	PowerEdge R760 server
Nodes per rack	16
Chassis	2.5" chassis with up to 24 NVMe Direct Drives, two CPUs
Chassis configuration	Riser configuration 3, half-length, two 2-channel full-height slots (Gen4), two 16-channel full-height slots (Gen5), and two 16-channel low-profile slots (Gen4)
Power supply	Dual, hot-plug, fully redundant (1+1) 1100 W power supplies
Processor	Intel Xeon Gold 6426Y 2.5 G, 16 C/32 T, 16 GT/s, 38 M
Memory capacity	512 GB (sixteen 32 GB RDIMM, 4800 MT/s, dual rank)
Internal RAID storage controllers	C30, no RAID for NVMe chassis
Disk—NVMe	24 6.4 TB Enterprise NVMe, mixed-use agnostic drive, U.2
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 (24 6.84 TB NVMe drives)
Front-end networking	Two 25 GbE (SFP28)

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

PowerScale infrastructure

Dell Technologies recommends the configuration in PowerScale configuration for storage in clusters using PowerScale for their primary lakehouse storage using HDFS.

Machine Function	Component
Model	PowerScale H7000 (hybrid)
Chassis	4U node
Nodes per chassis	Four
Node storage	Twenty 12 TB 3.5-inch four native sector size 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

The recommended configuration is sized for typical usage as lakehouse HDFS storage.

Two Ethernet network ports per node included for connection to the Cluster data network or a PowerScale storage network. Two additional network ports are included for connection to the PowerScale back-end network. These additional ports can be either InfiniBand QDR or 40 GbE, depending upon on-site preferences.

One PowerScale H7000 chassis supports four PowerScale H7000 nodes. This configuration provides approximately 720 TB of usable storage. At 85% utilization, 600 TB of HDFS storage is a good guideline for available storage per chassis.

This configuration assumes that the PowerScale nodes are primarily used for HDFS storage. If the PowerScale nodes are used for other storage applications or clusters, you must account for it in the overall cluster sizing. You can also use other PowerScale H7000 drive configurations.