Dell Validated Design for Analytics — Data Lakehouse

A Dell Technologies Overview

August 2022

H19234

White Paper

Abstract

This white paper describes the use of a Data Lakehouse to streamline and optimize data analytics. A Dell Technologies Validated Design for Analytics – Data Lakehouse is introduced, featuring Dell PowerEdge and PowerScale infrastructure and supporting Robio.io and Delta Lake software by Delta.io for data lakehouse analytics.

Dell Technologies Solutions

Dell Technologies Validated Design

ROBIN A Rakuten Symphony Company

intel.

D✷**LL**Technologies

Copyright

Contents

Introduction

Executive summary

Digital transformation has moved businesses from a mode of retrieving and organizing critical information in conventional data stores to a new goal: capturing and storing every bit that passes through the business. The number and diversity of data sources are constantly expanding. New horizons are recognized in data as a raw resource with potential for value creation, even if specific points of value cannot yet be discerned.

The data lake was created to enable this new raw resource retention. The data lake excels at collecting the greatest amount and widest diversity of information. This data can range from the highly structured like market data, financial records, or transactions to massive or unstructured data such as application logs, sensor feeds, or rich media.

The length of development cycles and the increasing number of applications required by organizations can both interfere with responsive, real-time exploitation of the growing data lake. Incredible insights that can help an organization quickly uncover opportunities and drive efficiencies are there in the data lake—but most remains locked away by either time or process constraints.

Enter the data lakehouse, which combines the best of a data warehouse and a data lake. It sifts through structured, semistructured, and unstructured data to feed responsive and real-time including business intelligence, analytics, marketing, AI, and machine learning applications. Productive, reliable data lakehouse implementations depend on the continuing evolution of the data center. It is one of the many pressures that promote well-tuned, modernized data-center infrastructure, with resources orchestrated by a modern, containerized application architecture.

Dell Technologies is well positioned to bring all necessary products, components, technology ecosystems, and services together to deliver the modern data center and benefits of a data lakehouse. Organizations can unleash the combined power of all their data, freeing data scientists and data engineers to create value rapidly and reliably from massive and diverse raw data resources.

Dell's end-to-end validated design is a complete solution that brings together the platforms (servers, storage, networking, and software), services, OPEX pricing options, and on-demand and self-service capabilities. The Dell Validated Design for Analytics - Data Lakehouse is a collaboration with Robin.io and includes Delta Lake technologies. It is an engineered, tested, and supported solution for addressing a new generation of analytics challenges that arise from extracting actionable data from massive data stores. It puts data scientists and data engineers in real-time control of the design and deployment of workloads while keeping IT in control of security and governance.

Document purpose

This document helps personnel involved in analytics, data-center modernization, or designing service offerings for data lakes, data warehouses, and data lakehouses to better serve their customers. Modernized infrastructure and data transformation initiatives help IT shops, analytics providers, and data engineers deploy and use data lakehouses. Data analytics becomes more powerful, capable, and relevant to the goals of the organization.

Audience

This document is intended for enterprises with data lakes or a data lake strategy interested in empowering their organizations to act more quickly, effectively, and efficiently on their data. Audience roles include:

- Data and application administrators
- IT decision-makers
- Data scientists
- Data engineers

A data lakehouse can assist more traditional analytics customers looking to modernize their data collection. It can also help analytics systems to get more value from their data or standardize their data for modern analytics workloads.

Business challenges

Market environment

The world has moved from one of selective, focused data collection, primarily driven by structured transactional data housed in data warehouses. The new model is one where every aspect of the enterprise, market, and communities can be captured and cataloged, regardless of its immediate value. The cost for data storage has plummeted and the ability to extract value in the future continues at a rapid pace. These events often lead to data being extracted and stored without any particular cognizance of its ultimate value.

Recent developments in workloads like machine learning and artificial intelligence, however, require such large amounts of potentially disparate data to derive their insight, and are rapidly revealing that value. But new types and sources of data being generated now, not to mention new types of applications, will create pressure to build new solutions to act on that data.

The rapid development of public cloud systems and services has complicated efforts to develop these new systems. Data repositories, query engines, and analysis tools in the cloud have been highly attractive for their ease of adoption and utility cost model. These capabilities continue to evolve. Organizations that need to take full advantage of new data models and quantities have continued to build and operate on-premises systems.

More organizations embrace digital transformation, data transformation and increase their reliance on analytics to help guide their actions. Data lakes have been growing in popularity increasingly deployed both on premises and in the cloud. The data warehouses of the past relied heavily on structured data and focused narrowly on business intelligence or decision support. As new data sources from logs, sensors, equipment, and industrial IoT became available, a different repository was required. This new breed of data is largely either unstructured or semistructured, making it more difficult for applications to consume it directly without some intermediate resolution layer. Data lakes are optimized for large-scale collection of both tabular and nontabular information, and they can store both raw and transformed data. This approach provides great application flexibility, and promoting large-scale analysis of diverse data from multiple sources for greater insights.

As more unstructured data is now being brought into analytics environments, a new organizational approach is needed to simplify the collection and utilization of this information. A more modern data center must be enabled to support a transformation in how data is collected, managed, and used. The potential complexity of this change and

these new environments creates adoption and management challenges for IT under pressure to implement this new paradigm reliably and rapidly. The data scientists and data engineers who are tasked to convert information into more and deeper insights are under pressure to monetize the data.

Data collection, storage, and analysis capabilities in the cloud entails both on-demand usage and nearly infinite capacity. These factors have driven substantial data center development. The modern data center must operate on a similar self-service, scalable basis. Doing so closes gaps in time and process between resources and the developers and expert users who need them to drive insights and value. It is a business imperative now to shorten development cycles and give greater control to knowledge workers and strategic implementers. This imperative must not cause the organization to lose control of access, security, data integrity, or costs.

Data analytics evolution

Over time several generations of broad data architectures have arisen to address large-scale data collection and analysis as the types of information collected. Their uses have grown exponentially, and diversified. See Figure 1.
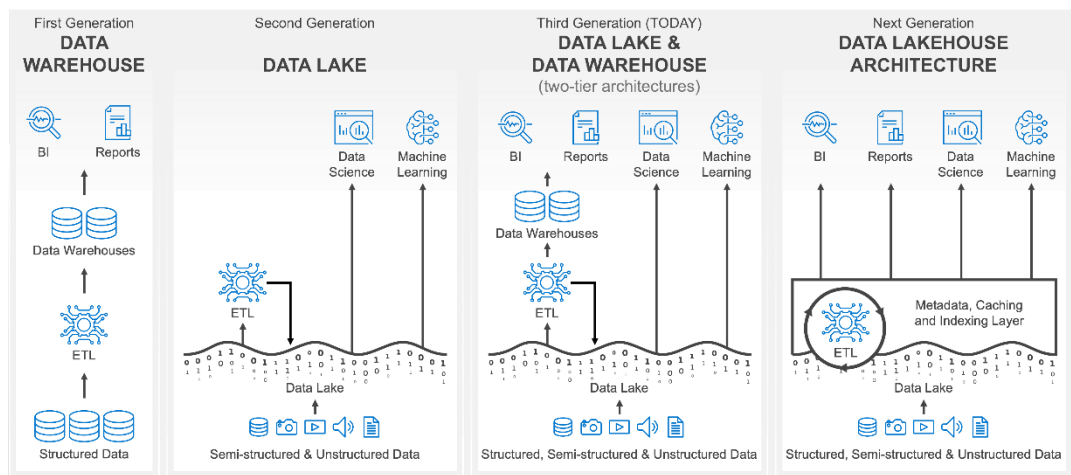
Figure 1.   The evolution of data analytics

Note: Figure 1 was adapted from "Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics," 11th Annual Conference on Innovative Data Systems Research, January 11–15, 2021, p. 2.

First generation

The first generation of broad data architecture designed to handle massive amounts of information were data warehouses. Data warehouses were optimized for business intelligence and decision support, dealing primarily with well-defined and structured data.

Using the foundational extract-transform-load (ETL) process, data could be presented to applications in the consistent manner that they require. ETL does have some degree of overhead and lacks a certain amount of flexibility. Businesses adapted to data warehouse requirements because of the valuable insights that a unified collection of many structured datasets could deliver.

Second generation

The second generation of architecture, the data lake, arose because data warehouses were not inherently dynamic. They could not easily scale to meet the changing needs of today's complex environments. Data lakes ingest both tabular and a wide variety of nontabular data such as objects, streams, hashes, and chains. Instead of having to apply a schema to data as it is written to the store, a new strategy of "schema on read" is used. Data lakes allow any type of data to be stored, regardless of format. The reading application enforces a schema at the time of extraction. The added benefit of the data being formatted for use rather than for storage is that of closing the gap between its collection and its value to the organization.

This paradigm was the origin of new applications for extracting data, like MapReduce and Spark. These applications draw information using novel queries from the new HDFS file system and accessing data in not only traditional formats but also as objects. Data lakes cannot support classic atomic, consistent, isolated, durable (ACID) transactions. They increase analytic capabilities with scalable metadata, session snapshot, unified batch and streaming, schema evolution, enforcement and audit transactions, and data manipulation language (DML) operations, for great flexibility.

Third generation

Today's third generation of architectures is the approach organizations are taking to enable both ACID, and dynamic and flexible, operations on increasing quantities and varieties of data. Data lakes have one downside to enabling the storage of any type of data. Quality, and governance cannot be applied during insert and must be moved downstream, closer to the point of extraction. Because of this, traditional tabular data that requires strict adherence to a schema is still kept in a data warehouse. Organizations therefore often use both a data warehouse and one or more data lakes. These two different systems with different purposes serving up either different data for the same subject areas or even the same data in different ways for the same subject areas. This strategy results in less reliability and the potential for stale data. For personnel working in analytics there became two sources to contend with, raising the overall complexity.

Next generation

The next generation of broad data architectures is the data lakehouse. It combines both a data lake and a data warehouse to avoid this duplication, complexity, and potential for analytical conflict. This design brings all the benefits of both kinds of systems into a single entity. Such a truly open format for data storage can provide either ACID, or dynamic and flexible operations, on the same data as required by the application. It also helps drive down costs while boosting scalability and adaptability at both the insertion and extraction of data. Management and governance can happen within the data lakehouse. A metadata, caching, and indexing layer watch all the data, the associated schemas (if required), and the state of the data, whether it is raw data or transformed results.

ETL becomes ELT

Accessing data lake information is a significant change from data warehouses, from past data access methods. The magnitude and diversity of a data lake often challenge businesses. The usual analytic process for data warehouse information is extract-transform-load (ETL) where data is queried, acted upon, and then loaded into the required destination for use. This method is applied to data lakes as well, with the added

requirement to comprehend and filter for amenable data structures. Many applications, though, must use other methods to access and use the data.
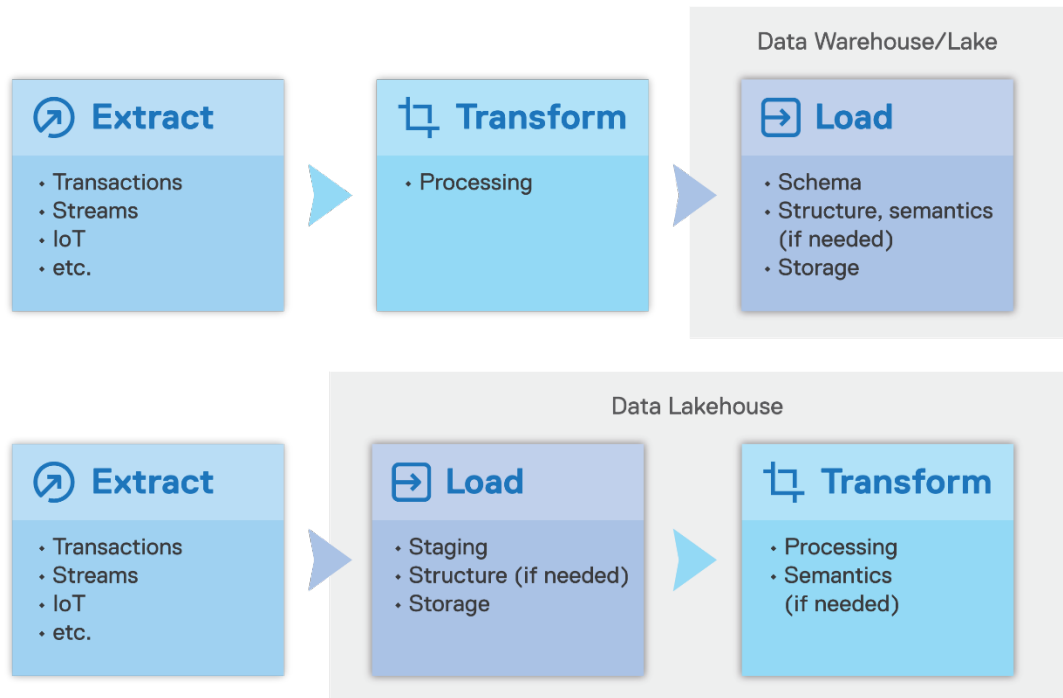
Figure 2.   Extract-load-transform (ELT) compared to the original extract-transform-load (ETL)

The ability of the data lakehouse to store transformed results, however, has changed the ETL paradigm, leading to a scenario where loading and transforming of data happens concurrently. More processing of the data is now happening within the data lake instead of at the point of the application, transitioning from ETL to extract-load-transform (ELT). The compute elements of the data lakehouse act upon the source data. The transformed results are stored within. The result is an acceleration of access for applications that need to tap into this rich, transformed source of information. See Figure 2.

Access and development are changing

A key issue for any data is timeliness and as organizations accelerate their decision timelines, the need for timely supporting data becomes more critical. A data warehouse-only strategy resulted in some data being deemed less valuable because of the time required to act on that data. Being able to accurately predict the weather in the next hour could be massively important to many enterprises. But if it takes two days to extract and act on that data, the value is virtually zero because the opportunity has passed.

The advent of cloud technology changed this paradigm by connecting developers, data scientists, and data engineers directly with both applications and data stores. But this paradigm often ceded control of critical elements like security away from the IT teams.

The time to develop analytical models increased the overall difficulty in accessing semistructured and unstructured information. It included time to create, validate mapping, apply other analytic functions, and for the required custom coding. The development of custom code for data access and having to handle change management is time consuming. The net result for organizations is a series of one-off solutions that have a lower overall applicability to the business.

With the rise of data lakes, the need for data scientists and data engineers became obvious. Data lakes were not as straightforward to access by LOBs as a data warehouse had been. But these data experts still need a layer of abstraction from the raw data, which is where a data lakehouse becomes valuable. It bridges the massive amounts of data with the applications that can consume and analyze the results.

Data experts today, like all users, also benefit greatly from the point-and-click, highly scalable, abstracted experience associated with the cloud. If they can spend less time and attention on process, maintenance, life-cycle operations, and systems management, the more they can spend on their core functions creating insights and value. This change has been profound in access and development, and enabling it is a critical component of any data analysis solution.

Organizations seeking a hybrid platform will need the same level of self-service and on-demand capabilities to which they are accustomed in the cloud world.

Solution overview

## A complete data lakehouse implementation

Based on the changing nature of data analytics and the need to access large amounts of disparate data, organizations need a new approach to data access. Data warehouses provided access to structured data, and data lakes handled the unstructured and semistructured data. At this stage, it was far too difficult to access both stores as an integrated entity. A data lakehouse combines the best of data warehouses and data lakes. It supports business intelligence and machine learning technologies in one platform that can store all types of data and provide it with a cloud-like, multiresource, self-service interface for data scientists. See Figure 3.
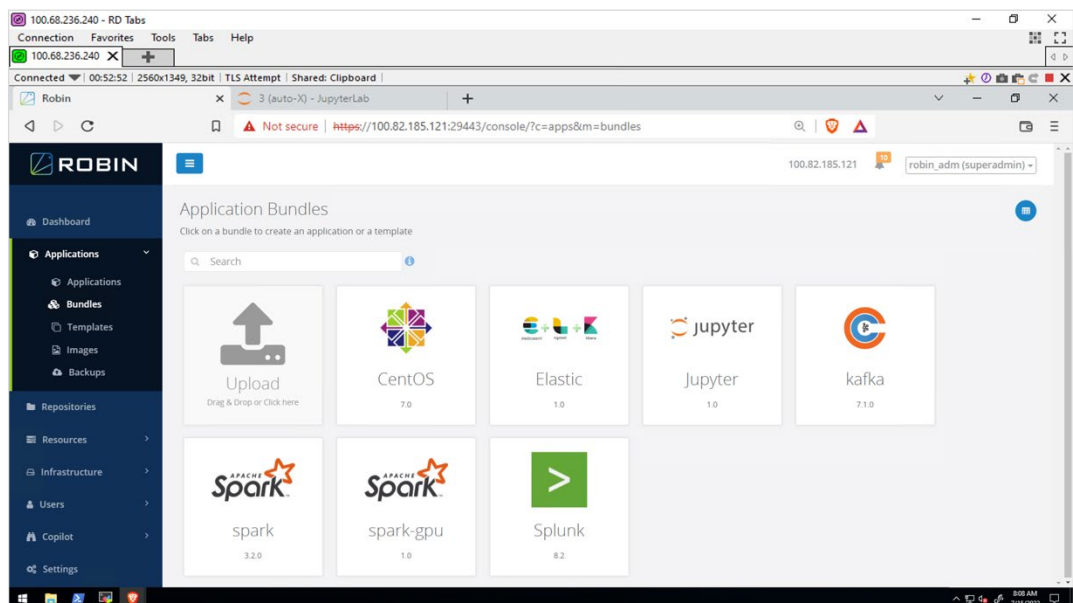
Figure 3.   A self-service data lakehouse interface for data scientists

This Validated Design offers an on-premises or co-located alternative to cloud-based data lakehouses. It includes Spark, Kafka, Delta Lake by Delta.io, Robin.io for hosting on a cloud-native, self-service, on-demand platform. It includes Intel® Xeon® -based Dell PowerEdge servers with optional Optane high-performance local storage, Dell

PowerSwitch networking, and Dell PowerScale or Dell ECS storage. See Figure 4. It delivers the ability to organize, orchestrate and automate compute, storage, and environments for the user from a single interface. That interface speeds access to results by removing many of the hurdles that data scientists and data engineers traditionally encounter when working with infrastructure.
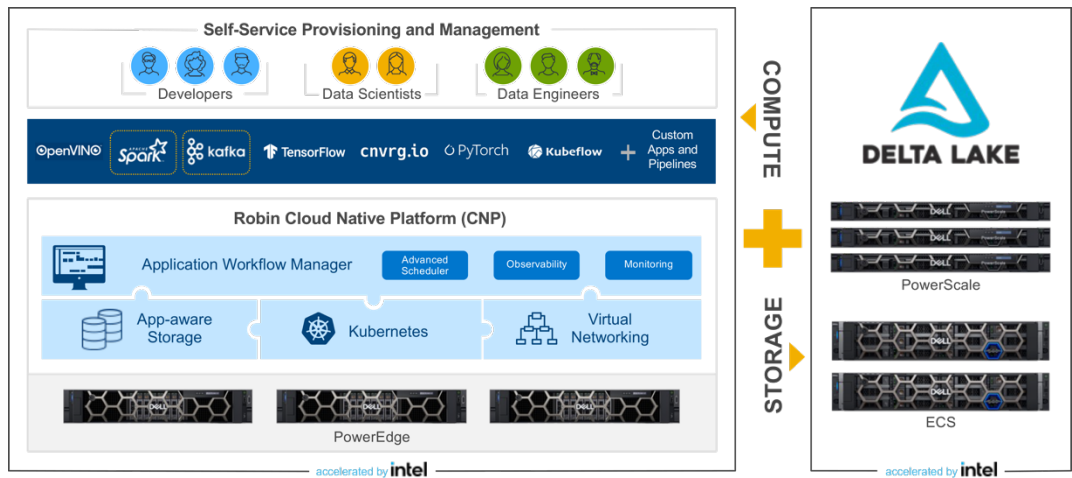
Figure 4.   The Dell Validated Design for Analytics – Data Lakehouse

This open data management platform combines the flexibility, cost-efficiency, and scale of data lakes with the data management and ACID transactions of data warehouses. The platform enables business intelligence and machine learning on structured, semistructured, and unstructured data from a single source.

Create value from data

The Dell Validated Design for Analytics – Data Lakehouse can run critical analytics projects featuring a faster time-to-insight cycle that results in better data quality and control. The net result for your organization is faster business intelligence and reporting.

Data engineers and data scientists can be empowered through the solution's self-service, on-demand, tools and frameworks that can either be run on-premises or in a co-located facility.

Faster interactive queries, coupled with better data availability, facilitates more informed decision-making. And data lakehouse operations like caching, indexing, and data compaction help to increase performance, enabling organizations to access and process the data to quickly drive outcomes.

With data lakehouse solutions from Dell Technologies, customers using the solution to enhance fraud detection reported up to a 15x acceleration in research value recognition. The results were hundreds of millions of dollars in cost avoidance for the business.

Simplify your data landscape

With a data lakehouse, your data management is simplified since, as stated above, structured, semistructured, and unstructured data can all be stored in the data lake. This design reduces the need for chasing, copying, or moving data between architectures. It gives data scientists and data engineers the ability to access all types of data from a single repository.

With support for ACID transactions, the Dell solution helps ensure consistency, even as multiple parties concurrently read, or write, data. Complexity and guesswork are

eliminated by leveraging Validated Design for Analytics – Data Lakehouse through Dell's thorough validation processes.

With solutions from Dell Technologies, customers report saving up to 12 hours per week through the automated reconciliation of data feeds. The tested and proven configuration delivers 18–20% faster configuration and integration with a 25% reduction in support.

Partner technology overview

Delta Lake by Delta.io

A Linux Foundation project since 2019, Delta Lake by Delta.io is an independent project controlled by a development community rather than any single technology vendor. The Dell Validated Design for Analytics – Data Lakehouse enables reliable deployment and operation of Delta Lake in the solution. More than 150 developers from over 50 different organizations associated with Delta Lake working on multiple storage repositories are all engaged daily to help push the program goals forward. Key Delta Lake capabilities include:

- ACID transactions for Spark which bring serializable isolation levels that ensure readers always see consistent data

- Scalable metadata handling, through Spark's distributed processing power, to easily handle all the metadata for petabyte-scale tables, with billions of files

- Streaming data ingest, batch historic backfill, and interactive queries that work by default with ease because Delta Lake tables are batch tables, a streaming source and sink

- Schemas that are automatically enforced, handling schema variations to prevent insertion of bad records during ingestion

- Data versioning to enable rollbacks, full historical audit trails, and reproducible machine learning experiments

- Merge, update, and delete operations that allow easy handling of complex use cases like change-data-capture, slowly changing dimension (SCD) operations and streaming upserts

Robin.io

The San Jose-based Robin.io (formerly Robin Systems) is a Rakuten Symphony company. Robin's Cloud Native Platform (CNP) is an advanced Kubernetes cloud platform built to provide an application-as-a-service and cloud-like experience, automating deployments and life-cycle operations of deployed applications.

Robin CNP provides an intuitive, declarative interface that reduces deployment complexity, timelines, and human error. The user inputs only what resources to include. The platform builds a reusable policy, models all the resource configurations, and autoconfigures them across the service's entire life cycle, including instantiation, start, stop, migrating, scaling, and deletion.

Robin.io enables a solution that is both self-service and on-demand for an on-premises or co-located environment. See Figure 5. Robin brings cloud-native agility to complex data-centric applications. It enables developers, data scientists, and data engineers to easily deploy any application pipeline in minutes from within their own storage repository with an app store styled experience.
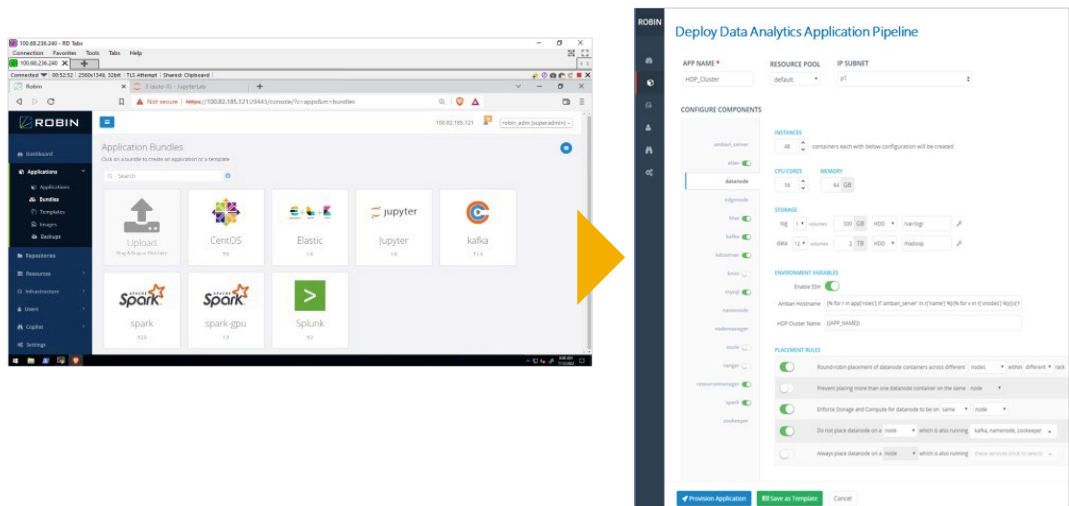
Figure 5.   App store like automated application delivery with Robin CNP

Robin's Application Workflow Manager brings end-to-end automation driven through either a one-click interface or driven completely by APIs, deploying entire application pipelines seamlessly. The enterprise-grade storage stack delivers all the application-aware features that users expect. These features include snapshots, clones, replication, backup, data rebalancing, tiering, thin provisioning, encryption, and compression.

The platform is designed to scale out as additional datasets are added to solutions. Servers can scale up as data needs increase and datasets expand, capturing an increasing amount of information. Through intent-based workload placement, Robin delivers guaranteed quality of service (QoS) through automation. It reports status through its QoS dashboard. With an easy, one-click cloning mechanism, Robin can speed time to production by cloning entire application stacks and data pipelines.

The Robin Cloud Native Platform offers integrated multitenancy to enable a shared cloud experience, with physical and logical separation between tenants. It includes integrated support for role-based access control to manage end-user credentials across each tenant and support for chargeback features to enable multiple departments and use cases.

The Apache Software Foundation

The Apache Software Foundation is an open-source, community-led foundation that is responsible for Apache Server, one of the most influential software components that drives both the Internet and enterprise computing. Two critical Apache components to the data lakehouse strategy are Kafka and Spark. Apache Kafka is a distributed event store and an event streaming platform that is designed to capture streams of data and events for use in analytics. Kafka is most efficient for use with real-time analysis solutions, relying on streams of events. Apache Spark is a multilanguage engine that can be used for data engineering, data science, and machine learning. A critical component of Spark is the ability to program clusters, providing parallelism and fault tolerance. Spark assists with the curation of data within the data lakehouse. It helps with pruning, cleaning, and maintaining the data, especially streamed content or batch processing. In the data lakehouse architecture, Apache Kafka interacts with the system through Spark.

Future technologies

The Dell Validated Design for Data Analytics is the platform of choice for data analytics moving forward. Dell Technologies will continue to expand this solution and deliver new

ones as a continuous evolution of the platform to further enhance its functionality for customers.

Solution architecture

Overall framework

Dell has worked closely with its software partners to develop a solution that addresses the growing challenges to analytics caused by the architectural differences between data warehouses and data lakes. The Dell Validated Design for Analytics – Data Lakehouse addresses the needs for organizations already down the path of deploying data warehouses and data lakes today. It also addresses the needs of organizations that have not deployed yet but recognize the value that advanced analytics can deliver to their decision making.

The Dell Technologies strategy combines compute and storage into a seamless solution. It gives developers, data scientists, and data engineers access to a self-service, one-click deployment of advanced analytic pipelines, with no need for extensive infrastructure interaction. See Figure 6.
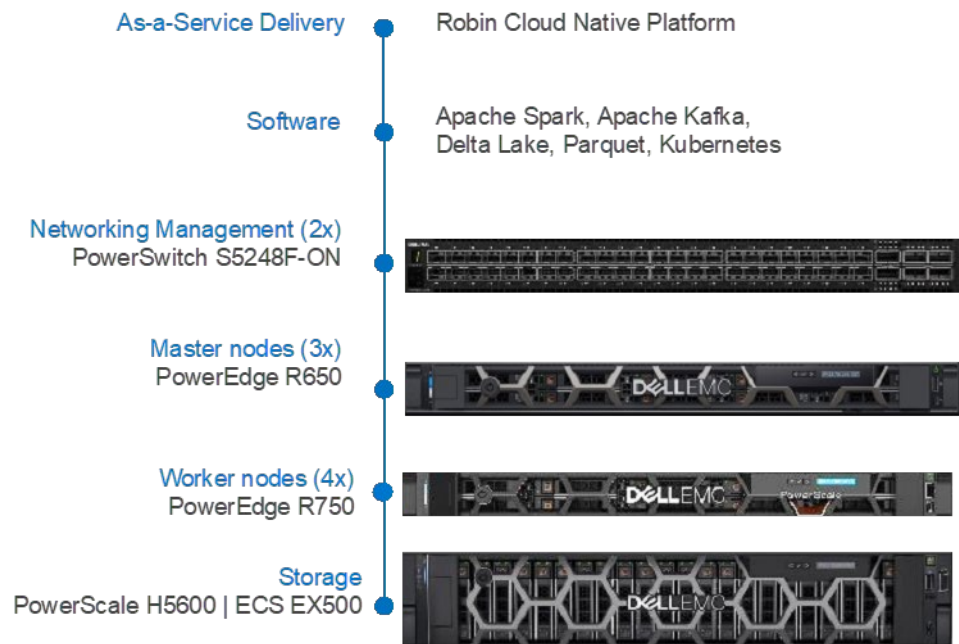
Figure 6.   Solution stack

This analytics solution is designed for deployment across a wide range of environments. The validation process assures that the solution provides flexibility for organizations to configure according to their needs.

System components

The Dell Validated Design for Data Analytics – Data Lakehouse draws on the strength of multiple Dell platforms. These platforms have been carefully chosen for their ability to handle compute-intensive workloads and scale to the demanding needs of data lakehouse environments.

- The **Dell PowerEdge R650** server is perfectly matched for compute-intensive workloads while also minimizing the data center footprint through it 1U form factor.

The PowerEdge R650 design enables business to easily scale while still handling challenging and emerging workloads. This dual-socket compute platform is perfectly matched to the management and control needs for the solution's control nodes. Control nodes handle the resource allocation, scheduling, system monitoring while also maintaining state across the cluster.

- The worker nodes handle all the "heavy lifting" for the solution. The choice of the **Dell PowerEdge R750** server easily handles the containerized environment, with the ability to change based on the workload demands and compute needs. The full featured enterprise server features a flexible 2U enclosure which provides the additional room for adding I/O bandwidth and storage capacity. Extensive GPU and storage support bring the ultimate in flexibility for workloads.

- The **Dell PowerSwitch S5248F-ON** delivers the top of rack switching. It is a 25 GbE/100 GbE open networking switch that provides state-of-the-art, high-density switching. The open networking capability provides extra flexibility for changing network configurations. The software defined networking enables a communications fabric that can change and adapt to the containerized compute cluster. The fabric optimizes communications paths with the compute demands as workloads are deployed and shift across the worker nodes.

- The **Dell PowerScale H5600** hybrid storage platform delivers a versatile yet simple scale-out storage architecture accessing massive amounts of data. With scalability of up to 1.28 PB per chassis and up to 8 GB/s of data bandwidth, the PowerScale H5600 can support demanding data lakehouse environments. The OneFS powered scalable storage brings the ability to speed access to massive unstructured data stores that can help feed data-hungry applications and analytics. The H5600 achieves up to 80% storage utilization compared with the 50% of traditional storage solutions. It is better suited for the demands of data lakehouses and continually changing underlying compute models.

Some data lakehouse include workloads that demand both unstructured data storage and object-based storage. The **Dell ECS EX500** delivers the perfect blend of economy and density for modern applications or deep archive environments. With scalability of up to 16 nodes, the ECS EX500 can handle up to 6,144 TB of unstructured storage per rack. Dell ECS enables enterprise-grade cloud scale storage for unstructured (object and file) storage while maintaining control, in a private cloud environment. The software-defined storage is layered to help enable a limitless scalability while maintaining the abstraction that promotes high availability.

System configuration

For the compute elements running the Robin Cloud Native Platform, the solution allows for the overlay of management services with compute and storage. See Table 1.

Table 1.   System elements

Component	Configuration
Control nodes	Minimum 3x Dell PowerEdge R650
Worker nodes	Minimum 4x Dell PowerEdge R750 NVIDIA GPU (optional)
CPU	Intel® Xeon® Scalable

Component	Configuration
Networking	Minimum 2x Dell PowerSwitch S5248F-ON
Storage	Minimum 3x Dell PowerScale H5600 *OR* Minimum 5x Dell ECS EX500 Enterprise Object Storage
Software	Apache Spark Apache Kafka Delta Lake Jupyter Notebook Parquet
Robin Cloud Native Platform	Robin Cloud Native Platform

The control nodes are validated as PowerEdge R650 servers and the worker nodes as PowerEdge R750 servers. Robin Cloud Native Storage (CNS) manages the NVMe or SSD storage disks. If GPU compute is required, PowerEdge R750 should be deployed as those platforms are better able to handle the additional power and cooling required for NVIDIA GPU accelerators. From a configuration perspective, dual-socket servers with at least 50 cores total are preferred. Typically, 32-core CPUs (64 total cores per node) are optimal. Memory configurations are configured starting at 512 GB.

For the storage elements, Dell PowerScale or Dell ECS should be deployed to handle the S3 storage and object store. The exact sizes and configurations of the storage will vary greatly from deployment to deployment based on both the size and types of information being stored. Cloud-native storage from Robin.io is deployed across the compute nodes for use by the Spark, Kafka, TensorFlow, PyTorch, and other runtimes.

For the Dell PowerScale configuration, a three-node H5600 configuration with roughly 500 TB per node is recommended. For Dell ECS deployments, a five-node EX500 configuration with 960 TB is the appropriate starting point, again based on the storage requirements.

Dell PowerSwitch S5248F-ON is recommended for the networking component of the solution.

Dell Technologies helps you every step of the way

The Dell Technologies Customer Solutions Center

The Dell Technologies Customer Solution Center (CSC) is an important supporting organization for your analytics implementation. The CSC uses the joint solution and many other solutions to accelerate achievement of your goals and realize your digital future:

- **Proof of Concept**—Validate that your preferred solution meets your needs with a custom Proof of Concept. Dell Technologies solution architects enable practical, firsthand implementation based on your test cases.

- **Design Session**—Collaborate with Dell Technologies experts to design a solution framework. Brainstorm with our experts to explore your current IT environment, your future objectives, and potential solutions.

- **Technical Deep Dive**—Dive into the technical solution details that you are considering for your organization. Learn from live product demonstrations and solution-focused discussions with Dell Technologies subject matter experts.

Contact your Dell Technologies Sales Representative today to schedule a customized briefing or solutions engagement for this or any other Dell Technologies Validated Design.

Deployment and support

Dell Technologies can provide a broad set of capabilities for implementing and maintaining solutions, linking people, processes, and technology to accelerate innovation and enable optimal business outcomes.

- Consulting Services help you create a competitive advantage for your business. Our expert consultants work with companies at all stages of data analytics. They can help you plan, implement, and optimize solutions that enable you to unlock your data capital and support advanced techniques, such as AI, ML, and DL.

- Deployment Services help you streamline complexity and bring new IT investments online as quickly as possible. Leverage our 30 plus years of experience for efficient and reliable solution deployment to accelerate adoption and return on investment (ROI) while freeing IT staff for more strategic work.

- Support Services driven by AI and DL will change the way you think about support with smart, groundbreaking technology backed by experts to help you maximize productivity, uptime, and convenience. Experience more than fast problem resolution—our AI engine proactively detects and prevents issues before they impact performance. Select ProSupport Plus for a single point of contact for hardware support.

- Payment Solutions from Dell Financial Services help you maximize your IT budget and get the technology you need today. Our portfolio includes traditional leasing and financing options, and advanced flexible consumption products to let you leverage OPEX instead of CAPEX if that suits your business requirements.

- Dell Technologies APEX offers a simple approach that gives you a wide range of consumption models, payment solutions, and services so you can optimize for various factors while realizing more predictable outcomes.

- Residency Services provide the expertise needed to drive effective IT transformation and keep IT infrastructure running at its peak. Resident experts work tirelessly to address challenges and requirements, with the ability to adjust as priorities shift.

Conclusion

For the next generation of data architectures

Data analytics is changing, and the value to organizations cannot be understated. Data warehouses and data lakes were both important steps forward in the enablement of consuming large pools of data. But both had limitations that prevented data scientists and data engineers from truly exploiting the information and extracting value from it. A data lakehouse enables organizations to better tap all types of data, be it structured, semistructured, or unstructured, to feed machine learning and artificial intelligence applications. The Dell Validated Design for Analytics – Data Lakehouse brings the power

of a data lakehouse with the confidence of a Dell-validated solution. It takes the guesswork out of designing and deploying the advanced analytics engine.

We value your feedback

Dell Technologies and the authors of this document welcome your feedback on the solution and the solution documentation. Contact the Dell Technologies Solutions team by email.

This white paper may contain language from third-party content that is not under Dell Technologies' control and is not consistent with current guidelines for Dell Technologies' own content. When such third-party content is updated by the relevant third parties, this white paper will be revised accordingly.

Terminology

The following table provides definitions for some of the terms that are used in this document.

Table 2.   Terminology

Term	Definition
ETL	Extract, transform, load
ELT	Extract, load, transform

References

Dell Technologies documentation

The following Dell Technologies documentation provides additional and relevant information. Access to these documents depends on your login credentials. If you do not have access to a document, contact your Dell Technologies representative.

- *Info Hub for Artificial Intelligence*

- *Info Hub for Data Analytics*

Conference on Innovative Data Systems (CIDR) documentation

Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics

Delta Late documentation

Delta Lake Community Web Page

Robin.io documentation

Robin Cloud Native Platform