Solution Brief

Enterprise, CoSP

Dell Taps Intel and Rakuten Symphony for Private Cloud Analytics

Dell Validated Design for Analytics enables easy setup of data lakehouse solutions using Rakuten Symphony Symcloud™ platform and Intel-powered Dell servers and storage

Authors

Jay Limbasiya
Dell

Mike King
Dell

Mehran Hadipour
Rakuten Symphony

Jerry Huang
Intel

Enterprise computing is going through a new wave of increased interest in hybrid cloud computing as business leaders assess the practicality and cost effectiveness of keeping certain workloads in the public cloud versus reverse migrating them to private cloud implementations.

The cost benefit of the cloud is great for dynamic workloads, due to the ability to provide scalable storage and compute resources in the public cloud. But analytics workloads have large data models and ingest large amounts of data, which can be very costly for a workload that doesn't have the need for rapid scaling. Also, many cloud providers do not guarantee processing performance and some have limits on compute power which is a challenge for compute-intensive applications like data analytics.

Cyber security protection is another cloud sticking point for certain workloads. Not all cyber security services are provided by the cloud provider and even what they do provide might not meet the needs of a particular workload. This adds confusion and an additional cost to the service when the enterprise has to provide its own security infrastructure.

One answer to these challenges is to repatriate the workloads into private cloud servers or to adopt a hybrid cloud strategy for these workloads with data being split between the private and public clouds. A 2022 white paper from IDC[1] shows an increase in interest in repatriating workloads into private clouds with over 70% of respondents saying they will be moving workloads into private cloud or non-cloud infrastructure over the next two years (see Figure 1).

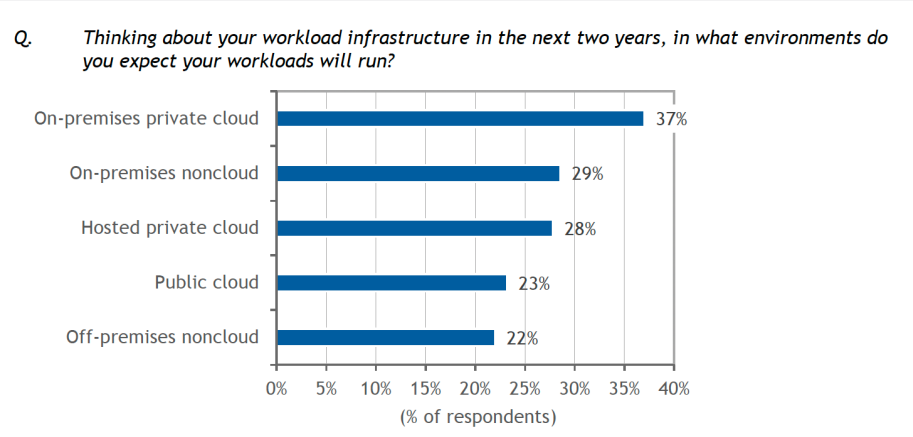

Figure 1. Source: IDC's 1H21 Servers and Storage Workloads Survey, August 2021. Survey includes 7,487 workloads across 2,325 respondents. Multiple responses were allowed.

Bringing Data Analytics to Private Clouds

One workload that makes financial sense to repatriate is compute intensive big data analytics, including workloads such as Hadoop®, data lakes, data lakehouses, Splunk®, busy NoSQL databases and artificial intelligence / machine language (AI/ML) workloads.

Intel Network Builders ecosystem partners Dell and Rakuten Symphony have contributed to the Dell Validated Design for Analytics — Data Lakehouse to provide a powerful solution that is also easy to deploy and operate.

A data lakehouse combines the capabilities of a data lake and a data warehouse. Data lakes can collect and store very diverse data types that are structured, semi-structured and unstructured. Unstructured data is a challenge because it can be from any source – sensors, rich media, geo-spatial data, audio, weather data and more. Data warehouses provide the data analytics on relational data that has already been processed.

The data lakehouse is a tool to process very diverse data types from data lakes by layering on the analytics functionality of the data warehouse. The data lakehouse can sort through structured and unstructured data in real time providing processed and relevant data for business intelligence, analytics and marketing applications.

The Dell Validated Design for Analytics — Data Lakehouse simplifies the creation of private servers for data lakehouse applications leveraging the cloud-native infrastructure and storage capabilities of the Rakuten Symphony Symcloud™ infrastructure running on Intel® architecture based servers and storage from Dell.

The pretested solution reduces the significant effort that faces data engineers and scientists when they deploy operational analytics or AI stacks. Many common AI and analytics applications have been pretested with Symcloud™ Kubernetes® containerization infrastructure. The blueprint also simplifies scaling of the solution. And finally, the blueprint has automation that make ongoing operations easy.

Rakuten Symphony Symcloud™ Platform

The Rakuten Symphony Symcloud™ platform is an industry-leading Kubernetes platform that is optimized for running storage and network intensive applications such as data lakehouses. The platform has the ability to automate complex applications in minutes and deliver cloud-native agility and scalability for complex data-centric applications – perfect for data scientists setting up their own private cloud.

Symcloud™ platform provides an Application Workflow Manager layer on top of Kubernetes that can onboard new applications via its one-click interface or via an API. The software can deploy complete application pipelines seamlessly. Once the software is onboarded, the Application Workflow Manager offers a slew of day 2 management tools including an advanced scheduler, application observability and monitoring.

Symcloud™ platform also provides app-aware storage that is important for managing the variety of data types that come in a data lake environment. Symcloud™ storage is an enterprise-grade storage stack that aggregates all of the available storage and distributes it to all of the applications.

It is also able to provision the storage to accommodate the data type. Symcloud™ platform also has a full slate of application-aware storage features including snapshots, clones, replication, backup, data rebalancing, tiering, thin provisioning, encryption, and compression.

Symcloud™ platform simplifies "as a service," on demand application delivery models that offers TCO benefits as well as time to value advantages and operating expense consumption models that operate much like the cloud.

Hardware: Dell PowerEdge™ Servers and PowerScale™ Storage Family

The foundation of the Dell Validated Design for Analytics —Data Lakehouse are Dell PowerEdge™ R760 servers that use 4th Gen Intel® Xeon® Scalable processors.

The versatile Dell PowerEdge™ R760 products are two socket, 2RU-high, air-cooled servers supporting up to 52-core (104-thread) Intel® Xeon® processors. The servers support up to 32 DDR5 RDIMMS of very fast RAM (running at up to 4800 MT/sec) for accelerated in-memory workloads.

The server family features eight universal drive slots for peripherals. Although built in storage is not applicable to the data lakehouse application, the Dell PowerEdge™ R760 servers can accept up to a maximum of 216 TB of storage. Support for graphic processor unit (GPU) cards - up to two double-wide or six single-wide GPUs – is built into the servers to finish workload processing faster.

Dell PowerScale™ H7000 Storage

For storage, the Dell Validated Design for Analytics —Data Lakehouse specifies PowerScale™ H7000 secure, scale-out network attached storage (NAS) drives. The PowerScale™ H7000 is a hybrid storage platform that offers up to 1.28 petabytes of scale-out storage per cluster with data bandwidth of up to 8 Gbps for fast access to massive unstructured data stores that can help feed data-hungry applications and analytics.

The PowerScale™ H7000 offers a choice of all-flash, or hybrid drives with the support for multi-cloud and native cloud deployments. The PowerScale™ H7000 has the performance, capacity and security features needed for deployment at the network edge, in the enterprise data center or in the cloud. The PowerScale™ H7000 handles any unstructured, structured or semi-structured data with the ability to scale performance, capacity and efficiency.

The PowerScale™ H7000 features a software defined architecture based on the PowerScale™ OneFS operating system that enables up to 80% storage utilization – which is above the industry average.

4th Gen Intel® Xeon® Scalable Processors

4th Gen Intel® Xeon® Scalable processors offer high-throughput and low-latency and are engineered for on-prem or cloud deployments. The processor family's architecture combines high-performance processor cores with up to eight built-in accelerators[2] for maximum performance efficiency.

These processors offer up to 80 lanes of PCIe 5.0 connectivity and support Compute Express Link (CXL), a cache-coherent interconnect for processors, memory expansion, and accelerators.

Integration of accelerators into the processor redefines CPU architecture and provides a more efficient way to achieve higher performance than relying solely on increasing the CPU core count for workload processing.

With all-new accelerated matrix multiply operations, 4th Gen Intel® Xeon® Scalable processors have exceptional AI training and inference performance. Other seamlessly integrated accelerators speed up data movement and compression for faster networking, boost query throughput for more responsive analytics, and offload scheduling and queue management to dynamically balance loads across multiple cores.

Putting It All Together

The use of a data lakehouse simplifies an organization's data management activities through its support of all data types. The Dell validated design is designed to simplify the deployment of data lakehouses on a private server by pre-integrating a number of the applications that are important to data scientists for analytics and data lakehouse operations.

These include Apache Spark™, Apache Kafka®, and Delta Lake open-source storage layer for data lakes. A full range of AI tools such as OpenVINO™, TensorFlow™, PyTorch™ and others have also been integrated into the solution.

Figure 2 presents the entire solution. Developers, data scientists, data engineers and other users can use the app store functionality built into Symcloud™ platform to spin up applications along with the container infrastructure without training or deep experience with server deployments.

The storage functionality built into Symcloud™ platform allows users to access attached storage so that it can provide unified data stores to an application.

The Validated Design supports ACID transactions (those that meet atomicity, consistency, isolation, and durability criteria) for guaranteed data validity.

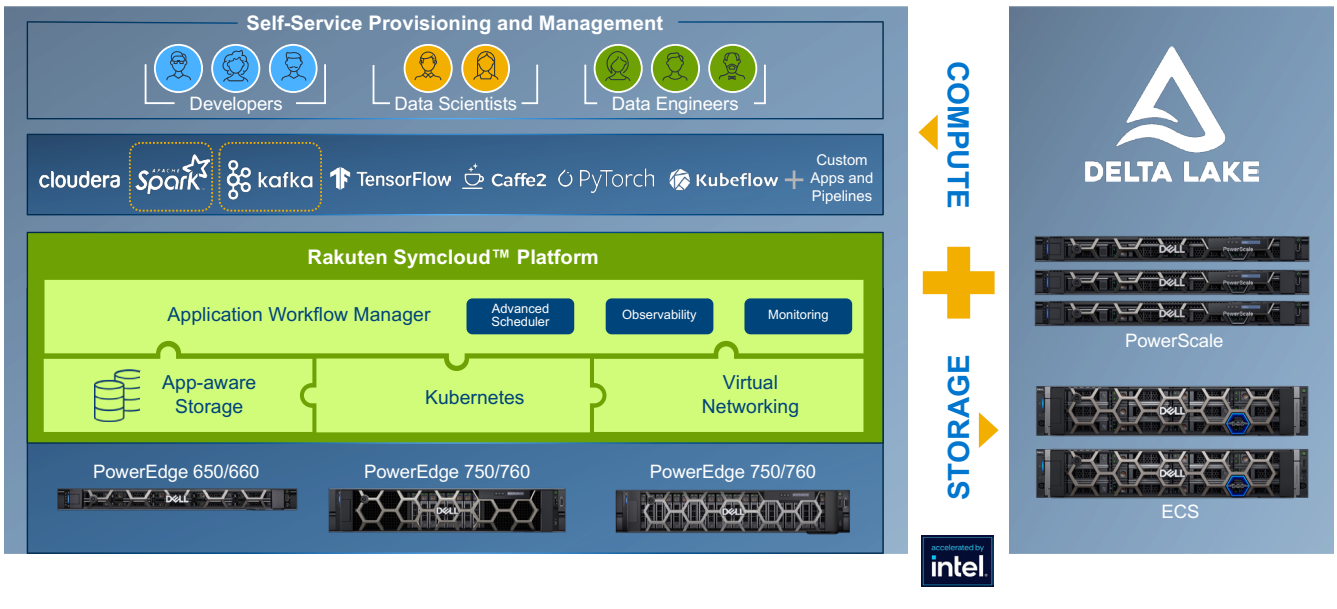

Figure 2. Graphical depiction of The Dell Validated Design for Analytics - Data Lakehouse.

Conclusion

Developing private cloud servers for data analytics applications and cloud optimization of these workloads is an increasing trend. But it's a significant effort for data scientists to develop their operational AI or analytics stacks – taking months to set up clusters and even more time to scale and manage. The Dell Validated Design for Analytics —Data Lakehouse reduces that time investment with applications that are integrated with the Symcloud™ cloud native infrastructure and the Intel architecture CPU-based Dell compute and storage platforms.

This blueprint provides a solution to a significant challenge facing enterprises today and will help them to get more out of their data potentially for less cost.

Learn More

Dell PowerEdge™ R760 servers

Dell PowerScale™ storage

Rakuten Symphony Symcloud™ platform

4th Gen Intel® Xeon® Scalable processors

Intel® Network Builders ecosystem

[1] https://www.supermicro.com/white_paper/IDC_On-Prem_Cloud_Success_Stories.pdf
[2] https://www.intel.com/content/www/us/en/products/details/processors/xeon/scalable.html

1023/DC/H09/PDF ♻Please Recycle 357182-001US