Lenovo Big Data Solutions for

Cloudera Data Platform

Enabling analytics and machine learning on growing business data

The Big Data Challenge

Big data is more than a challenge. It is an opportunity to find new insights in data to make your business more agile and to answer questions that were previously beyond reach. To open the door to a world of possibilities Cloudera employs the latest big data technologies to address critical business value drivers – growing business, connecting products and services, and protecting business.

Apache Hadoop and Apache Spark are open source software frameworks that are used to reliably manage and analyze large volumes of structured and unstructured data. Cloudera enhances this technology to withstand the demands of your enterprise, adding management, security, governance, and analytics features. The result is that you get an enterprise-ready solution for complex, large-scale analytics.

The Lenovo Big Data Solutions for Cloudera Data Platform (CDP) provide a predefined and optimized hardware infrastructure for CDP Private Cloud, a hybrid cloud version of CDP that seamlessly connects on-premises environments to public clouds with consistent, built-in security and governance. These solutions enable analysis of large data sets easily and quickly through a massively parallel processing environment, and provide exceptional reliability, scalability and flexibility. Entry through high-end configurations are supported along with the ability to easily scale as enterprise use of big data grows.

Cloudera Data Platform

CDP is an enterprise analytics and management platform, enabling ingestion, management, and delivery of any analytics workload from Edge to AI. It provides enterprise grade security and governance, and self-service access to integrated, multi-function analytics on centrally managed and secured business data. CDP allows you to meet the exponential demand for analytics and machine learning services with a petabyte-scale hybrid data architecture, delivering faster time to value and supporting critical workloads at scale.

CDP gives you complete visibility into all your data. The CDP control plane allows you to manage the data, infrastructure, analytics, and analytic workloads across hybrid and multi-cloud environments, providing consistent security and governance across the entire data lifecycle.

Highlights

- Deliver analytics and machine learning services to react faster to changing business requirements
- Meet the growing demand for analytics and machine learning services with a scalable data architecture
- Consistently and easily enforce security and governance policies across hybrid and multi-cloud deployments to ensure regulatory compliance
- Invest in a platform powered by open source, ensuring continual and rapid innovation to address evolving business requirements

One Data Platform. Many Applications.

Cloudera Data Platform is the world's most complete, tested, and popular distribution of Apache Hadoop and related projects. All of the packaging and integration work is done for you, and the entire solution is thoroughly tested and fully documented. By taking the guesswork out of building out your Hadoop deployment, Cloudera Data Platform gives you a streamlined path to success in solving real business problems with big data.

CDP provides a consistent experience across Public Cloud, Multi-Cloud, and Private Cloud deployments.

CDP Private Cloud provides a disaggregation of compute and storage, and allows independent scaling of compute and storage clusters. Through the use of containerized applications deployed on Kubernetes, CDP Private Cloud brings both agility and predictable performance to analytic applications. The three main benefits of CDP Private Cloud are:

- **Simplified multitenancy and isolation** - The containerized deployment of applications in CDP Private Cloud ensures that each application is sufficiently isolated and can run independently from others on the same Kubernetes infrastructure. Such a deployment also helps in independently upgrading applications based on your requirements. In addition, all these applications can share a common Data Lake instance.
- **Simplified deployment of applications** - CDP Private Cloud deployment ensures a much faster deployment of applications with a shared Data Lake compared to monolithic clusters where separate copies of security and governance data would be required for each separate application. In situations where you need to provision applications on an arbitrary basis, for example, to deploy test applications; CDP Private Cloud enables you to rapidly perform such deployments.
- **Better utilization of infrastructure** - CDP Private Cloud enables you to provision resources in real time when deploying applications. In addition, the ability to scale or suspend applications on a need basis in CDP Private Cloud ensures that your on-premises infrastructure is utilized optimally.

GPU-enabled Acceleration

Lenovo has partnered with NVIDIA and Cloudera to integrate NVIDIA RAPIDS accelerated data science libraries on Cloudera Data Platform, enabling GPU-accelerated Apache Spark 3 applications. Apache Spark included in CDP has been a workhorse for numerous data analytics tasks such as batch/real-time streaming, data warehouse and machine learning among others. Accelerating Spark with GPU-enabled computation is the next leap forward in helping enterprises achieve the goal of faster and better model development. Key technical features include faster and more accurate analytics and prediction from machine learning models. The ultimate business benefits are significantly improved price-to-performance as well as return-on-investment (ROI) by leveraging better data analytics.

Customer Spotlight

A utility needed a unified and more holistic view of their company data - both legacy database and unstructured data. They consolidated their data with Cloudera on Lenovo ThinkSystem SR650 servers, and are now implementing predictive analytics on power consumption utilization as well as fraud detection.

Powered by Lenovo

Lenovo Solutions for Cloudera Data Platform provide an optimized hardware infrastructure designed for high performance and scalability, handling the Big Data analytics and machine learning requirements of your business today and in the future.

Lenovo Solutions for Cloudera Data Platform utilize cost-efficient, industry standard x86 servers. Recommended models include:

- SR650 V3 - 2U, 2-socket server based on 4th Generation Intel® Xeon® Scalable processors. The SR650 V3 provides a GPU-rich platform, an abundance of DDR5, and optimized data transfer rates for big data workloads. The SR650 V2 based on 3rd Generation Intel Xeon Scalable processors is also supported.
- SR630 V3 - 1U, 2-socket server based on 4th Generation Intel® Xeon® Scalable processors. The SR630 V3 supports up to 60 core processors, DDR5 memory, PCIe Gen5 technology and NVMe storage for high performance. The SR630 V2 based on 3rd Generation Intel Xeon Scalable processors is also supported.
- SR665 V3 or SR655 V3 - 2U, 2-socket or 1-socket servers, respectively, based on 4th Generation AMD EPYC™ processors. The SR665 V3 and SR655 V3 provide up to 128 cores per processor, a GPU-rich platform, and an abundance of DDR5 memory and NVMe storage for big data workloads. The SR665 and SR655 based on 3rd Generation AMD EPYC processors is also supported.
- SR645 V3 or SR635 V3 - 1U, 2-socket or 1-socket servers, respectively, based on 4th Generation AMD EPYC processors. The SR645 V3 and SR635 V3 provide up to 128 cores per processor, DDR5 memory, NVMe storage and up to 3 single-width GPUs for high performance. The SR645 and SR635 based on 3rd Generation AMD EPYC processors are also supported.

All Lenovo ThinkSystem servers are high performance systems, consistently holding numerous world performance benchmarks. Engineered for always-on productivity, ThinkSystem servers are consistently ranked high in x86 server customer satisfaction and #1 in x86 server reliability[1].

ThinkSystem SR650 V3

ThinkSystem SR645 V3

Complementary Offerings

Connecting the clustered server environment in these solutions can be easily accomplished with network switches. The recommended offerings for these solutions are 10GbE switches.

Lenovo XClarity™ Administrator is a centralized resource management solution that is aimed at reducing complexity, speeding response, and enhancing the availability of Lenovo server systems and solutions. It captures industry-leading proactive platform alerts, enabling administrators to migrate workloads or replace failing components without incurring downtime.

Tying It All Together

In today's rapidly-changing technology environment, empowering your data center transformation isn't just a necessity—it's also a journey. Regardless of your current environment, Lenovo Services is a true business partner that will take you from where you are, to where you want to be. At every stage, you'll get our expertise and services to help you:

- **Drive Digital Transformation.** You'll get the best architectures suited to your unique needs, along with our industry insights, expert guidance, and hands-on experience.
- **Foster Innovation.** Free up your internal resources to focus on initiatives that grow your business.
- **Simplify Your Support Experience.** Gain a trusted partner who understands your systems and solutions to fully support and optimize your data center.

Lenovo solutions for Cloudera Data Platform provide flexibility, scalability and high performance at a cost-effective price

Why Lenovo

Lenovo is a leading provider of data center infrastructure solutions. We partner with you to identify, design, install and support the solution that best ensures your organization's needs are met throughout the IT lifecycle. Lenovo complements a portfolio of leading x86 infrastructure with a full range of storage, software, and comprehensive services that provides excellent performance, reliability, and security for your IT environment from the edge to the cloud.

For More Information

To learn more about Lenovo solutions for Cloudera Data Platform, contact your Lenovo sales representative or Business Partner or visit: www.lenovo.com/systems

Reference Architecture:

Lenovo Big Data Reference Design for Cloudera Data Platform on ThinkSystem Servers

[1] ITIC reliability study, https://lenovopress.com/lp1117-itic-reliability-study

BDGCL02SB04