Intel Accelerated Solutions
Solution Snapshot

Modernizing the Data Pipeline with Cloudera Data Platform and Intel® Technology

The Challenge

Businesses continue to struggle with gaining meaningful insights from their data. The explosion of data volume to the prevalence of data and analytics silos increases these challenges even more. These need to be addressed along with effectively managing and securing an organization's data to prevent costly privacy breaches.

Cloudera Overview + Benefits

As data grows and analytics workloads get more complex, analytics workloads need to be flexible and highly responsive. Cloudera Data Platform (CDP) is a hybrid data platform designed for unmatched freedom to choose—any cloud, any analytics, any data. CDP delivers faster and easier data management and data analytics for data anywhere, with optimal performance, scalability, and security. With CDP, you get all the advantages of CDP Private Cloud and CDP Public Cloud for faster time to value and increased control.

Cloudera Data Platform

Use Cases

Deploying a Private Cloud

CDP delivers an end-to-end analytics platform that supports hybrid and multi-cloud environments and eliminates costly and inefficient data silos.

Predictive Analytics and Machine Learning

End-to-end machine learning pipeline with Cloudera and open-source frameworks.

Modern Data Pipeline

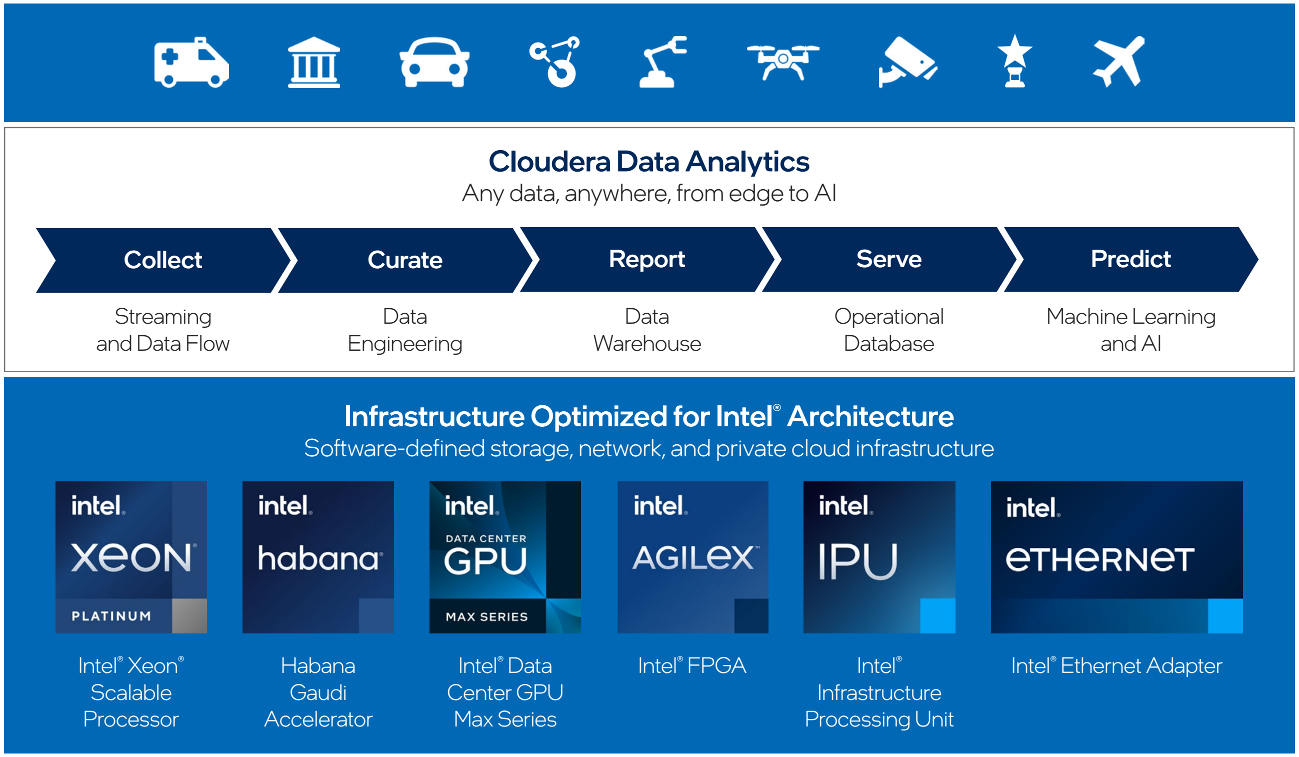
Apache Airflow-based pipeline provides orchestration for multi-step data pipelines in the cloud. The pipeline is based on a combination of Spark and Hive and generates curated datasets for downstream applications securely and efficiently.

Data Lake/Lakehouse Cloudera

Open Data Lakehouse helps organizations run quick analytics on all data—structured and unstructured—at a massive scale.

Applications and Solutions

Analytics-Powered Vertical and Horizontal Solutions



Cloudera Data Analytics
Any data, anywhere, from edge to AI

Collect	Curate	Report	Serve	Predict
Streaming and Data Flow	Data Engineering	Data Warehouse	Operational Database	Machine Learning and AI

Infrastructure Optimized for Intel® Architecture
Software-defined storage, network, and private cloud infrastructure

Intel® Xeon® Scalable Processor

Habana Gaudi Accelerator

Intel® Data Center GPU Max Series

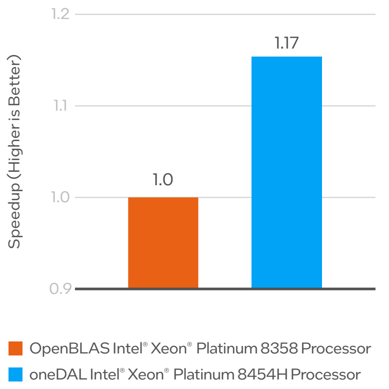
Intel® FPGA

Intel® Infrastructure Processing Unit

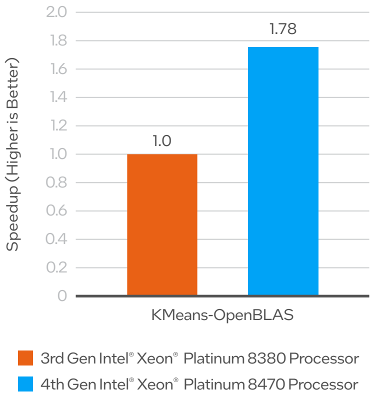
Intel® Ethernet Adapter

Proof Points

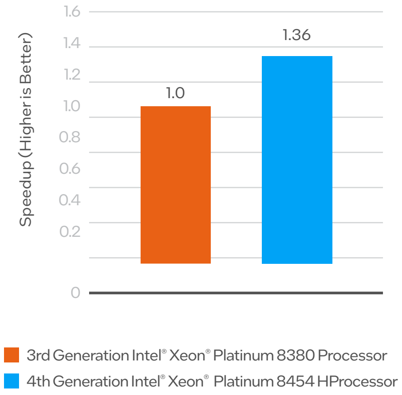
3rd Gen Intel® Xeon® Platinum 8358 Processor versus 4th Gen Intel® Xeon® Platinum 8454H Processor[1]



HiBench KMeans Performance Gain for 4th Gen Intel® Xeon® Scalable Processor[2]
(1 Master + 3 Worker Node)



Performance for Spark KMeans Clustering using OpenBLAS Library on Cloudera Private Cloud Base Cluster[3]



Up to 17% performance gain on 4th Gen Intel® Xeon® Platinum 8454H processor versus 3rd Gen Intel® Xeon® Platinum 8358 processor using Spark CDP Private Cloud Base with MLlib and Intel® oneAPI Data Analytics Library[1]

Up to 78% performance gain on Intel® Xeon® Platinum 8470 processor when compared to Spark to Intel® Xeon® Platinum 8380 processor for Spark KMeans algorithm using Spark CDP Private Cloud Base 7.1.7 and OpenBLAS library instances[2]

Up to 36% performance gain for KMeans Clustering (ML) using Spark CDP Private Cloud Base 7.1.7 on 4th Gen Intel® Xeon® Scalable processor versus 3rd Gen Intel® Xeon® Scalable processor[3]

Why Intel for Cloudera Data Platform?

Faster Machine Learning Algorithms

Cloudera Data Platform (CDP) uses Intel® accelerators, libraries, and instruction sets to enable several predictive analytics and machine learning workloads. These include Intel® Data Streaming Accelerator, Intel® In-Memory Analytics Accelerator (IAA), Intel® QuickAssist Technology (QAT), and Intel® Advanced Vector Extensions 512. There are also innovative, open-source technologies that are a part of the overall solution.

Reduce Complexity of Analytics Deployments

Cloudera and Intel joint solutions are workload-optimized to minimize the challenges of infrastructure evaluation and deployment. They accelerate deployment with validated hardware and software stacks. These solutions balance the needs of performance, scale, and availability to reduce deployment risk.

Easy Data Management

Cloudera Shared Data Experience (SDX), running on Intel® technologies, combine enterprise-grade centralized security, governance, and management capabilities with shared metadata and a data catalog to eliminate costly data silos, prevent lock-in to proprietary formats, and eradicate resource contention. SDX keeps your information secure by design with an integrated set of security and governance technologies to protect data. Security is further strengthened with Intel® Software Guard Extensions (SGX), which isolates sensitive data with hardware-based memory protections and Intel® Trusted Domain Extensions (TDX), used in virtualization environments where the entire virtual machine is isolated. Cloudera users can use these features to securely migrate their sensitive CDP Private Cloud workloads to CDP Public Cloud.

Solutions for Data-Intensive Workloads

Optimized Cloudera software applications combined with proven Intel® architecture are key to building highly performant data applications. They support all kinds of data ingested from cloud to edge and are robust enough to handle the challenges of modern workloads. The open and flexible platforms provide interoperability with existing tools, storage efficiency, and scale.

Want More Information?

To get more information on Intel and Cloudera joint solutions, visit www.cloudera.com/partners/solutions/intel.html or www.intel.com/cloudera

1. Tested by Intel as of 10/20/2022. Config 1: 5-node, 2x Intel® Xeon® Platinum 8358 Processor, 32 cores @ 2.6GHz , HT On Turbo ON Total Memory 512GB (16 slots/ 32GB/ 3200 MT/s), BIOS: SE5C6200.86B.0022.D64.2105220049(ucode: 0xd000363), Red Hat Enterprise Linux 8.6 (Ootpa), 4.18.0-372.9.1.el8.x86_64, 4xP5510 NVMe, Intel® Ethernet Controller E810-C for QSFP, Cloudera Runtime 7.1.7-1.cdh7.1.7.p1000.24102687, Spark 3.1.0, HiBench Kmeans, OpenBLAS 0.3.15

   Config 2: 5-node, 2x Intel® Xeon® Platinum 8454H Processor, 32 cores @ 2.1 GHz, HT On Turbo ON Total Memory 512GB (16 slots/ 32GB/ 4800 MT/s), BIOS: 3A05 (0x2b000070), Red Hat Enterprise Linux 8.6 (Ootpa), 4.18.0-372.9.1.el8.x86_64, 4xP5510 NVMe, Intel® Ethernet Controller E810-C for QSFP, Cloudera Runtime 7.1.7-1.cdh7.1.7.p1000.24102687, Spark 3.1.0, HiBench Kmeans, OpenBLAS 0.3.15.
2. Tested by Intel as of 2/20/22. Config 1: 4-node, 2x Intel® Xeon® Platinum 8380 Processor, 40 cores, HT On, Turbo On, Total Memory 512GB (16x32GB DDR5 3200 MT/s), BIOS 1.8.2, microcode 0xd00037b ,<OS Version 2x NetXtreme BCM5720 Gigabit Ethernet PCIe, 2x Ethernet Controller E810-C for QSFP, 1x Ethernet interface, 6x 2.9T Solidigm NVMe P5600 MU U.2 3.2TB, 1x 1.5T Dell Ent NVMe v2 AGN FIPS MU U.2 1.6TB, Red Hat Enterprise Linux 8.6 (Ootpa), 4.18.0-372.9.1.el8.x86_64 , Hibench-7.1.1, 3.1.1, CDH- 7.1.7, Netlib-java 1.1, OpenBLAS 0.3.15, Elapsed Time – 541 seconds

   Config 2: 4-node, 2x Intel® Xeon® Platinum 8470 Processor, 52 cores, HT On, Turbo On, Total Memory 512GB (16x32GB DDR5 4800 MT/s [4800 MT/s]), BIOS 0.2.27, microcode 0xab000080, 2x NetXtreme BCM5720 Gigabit Ethernet PCIe, 2x Ethernet Controller E810-C for QSFP, 1x Ethernet interface, 6x 2.9T Solidigm NVMe P5600 MU U.2 3.2TB, 1x 1.5T Dell Ent NVMe v2 AGN FIPS MU U.2 1.6TB, Red Hat Enterprise Linux 8.6 (Ootpa), 4.18.0-372.9.1.el8.x86_64 , Hibench-7.1.1, 3.1.1, CDH- 7.1.7, Netlib-java 1.1, OpenBLAS 0.3.15, Elapsed Time – 303 seconds
3. Tested by Intel as of 10/20/2022. Config 1: 5-node, 2x Intel® Xeon® Platinum 8358 Processor, 32 cores @ 2.6GHz, HT On Turbo ON Total Memory 512GB (16 slots/ 32GB/ 3200 MT/s), BIOS: SE5C6200.86B.0022.D64.2105220049(ucode: 0xd000363), Red Hat Enterprise Linux 8.6 (Ootpa), 4.18.0-372.9.1.el8.x86_64, 4xP5510 NVMe, Intel® Ethernet Controller E810-C for QSFP, Cloudera Runtime 7.1.7-1.cdh7.1.7.p1000.24102687, Spark 3.1.0, HiBench Kmeans, OpenBLAS 0.3.15

   Config 2: 5-node, 2x Intel® Xeon® Platinum 8454H Processor, 32 cores @ 2.1 GHz, HT On Turbo ON Total Memory 512GB (16 slots/ 32GB/ 4800 MT/s), BIOS: 3A05 (0x2b000070), Red Hat Enterprise Linux 8.6 (Ootpa), 4.18.0-372.9.1.el8.x86_64, 4xP5510 NVMe, Intel® Ethernet Controller E810-C for QSFP, Cloudera Runtime 7.1.7-1.cdh7.1.7.p1000.24102687, Spark 3.1.0, HiBench Kmeans, OpenBLAS 0.3.15.

Performance varies by use, configuration, and other factors. Learn more at www.Intel.com/PerformanceIndex

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software, or service activation.