Solution Brief

Cloud Workloads

Benchmarking Microsoft Azure Databricks on Intel®-Optimized Instances

Modern enterprises can take advantage of optimized Intel instances in Azure to help improve performance and lower total cost of ownership (TCO).

Authors

Swastik Chakraborty
swastik.chakraborty@intel.com

Lakshman Chari
lakshman.chari@intel.com

Summary

As enterprises move more artificial intelligence (AI) workloads to cloud platforms such as Microsoft Azure, knowing what's powering those workloads becomes crucial. This solution brief describes how Azure Databricks, a unified data-analytics platform, can help enterprises manage more data, and why it is important to run Databricks workloads on Intel hardware. The solution brief then compares Azure Databricks data-pipeline performance between two generations of Intel® Xeon® Platinum processors.

Today's enterprises generate, collect, store, and analyze large amounts of diverse data, both structured and unstructured. To help manage this data, enterprises are moving their data workloads to cloud-based machine learning (ML) and artificial intelligence (AI) platforms. Doing so also lets organizations harness their data to enhance business functions and capabilities.

These workloads can take advantage of cloud-specific AI and data-pipeline products to accelerate complex ML models, and to capture large datasets. Moving ML and AI workloads to the cloud can help enterprises scale as their needs change. However, it's important for architects and software developers to understand the underlying cloud hardware powering their workloads.

To enable faster data pipelines in Azure Databricks, enterprises can choose Azure virtual machines (VMs) that are based on 2nd Generation Intel Xeon Platinum processors. This paper explores how moving from 1st Generation Intel Xeon Platinum processors to 2nd Generation Intel Xeon Platinum processor–optimized cloud instances helps accelerate the processing times for data science and engineering workloads on Azure Databricks. It also explores how doing so can help reduce total cost of ownership (TCO).

Azure Databricks enables efficient data pipelining

For decades, enterprises have relied on ever-increasing volumes of data to manage their businesses. Common data-management and analysis solutions include data warehouses that are tied to decision-support systems. These warehouses are repositories that ingest data from multiple sources, such as operational data from enterprise resource planning (ERP) systems or customer sales data from customer relationship management (CRM) systems.

Data warehouses and decision support systems typically rely on rigidly structured relational databases that were never designed for high-velocity semi-structured or unstructured data. Semi-structured data can include Extensible Markup Language (XML), JavaScript Object Notation (JSON) documents, and e-mails. Unstructured data can include digital video feeds and raw sensor data from Internet of Things (IoT) devices.

Data warehouses and decision support systems still have their place in the enterprise. However, the data landscape is changing as these never-ending streams of high-velocity, highly varied data become more common.

It can be a challenge for any size of organization to ingest, store, and analyze high-velocity data. But as data-storage capacity and CPU processing power have increased, organizations of all sizes have begun to see the potential of capturing and analyzing high-velocity, highly varied data.

Beginning in the early 2010s, the concept of data lakes emerged. Data lakes are systems capable of storing massive amounts of raw data in a variety of structured, semi-structured, and unstructured formats. As with data warehouses, data lakes also have their drawbacks. While data lakes can store massive amounts of varied data, they don't support transactions, nor do they enforce data quality. Both features are found in data warehouses. As a result, enterprises often end up having to maintain both data warehouses and data lakes.

Azure Databricks is a data and AI platform, optimized for the Microsoft Azure cloud services platform. It combines the best elements of data lakes and data warehouses with the Databricks Lakehouse architecture. Azure Databricks:

- Enables big data pipelines
- Runs SQL queries to extract business intelligence (BI)
- Provides an integrated end-to-end ML environment

The Lakehouse architecture

A data Lakehouse combines the data structure and data-management features found in data warehouses with the low-cost, high-velocity storage capabilities of data lakes. The result is a new, open data management architecture. Built on top of popular open source projects, such as Apache Spark, Delta Lake, and MLflow, Azure Databricks supports Lakehouse architectures by:

- Providing a data-management layer for fast, direct access to open data formats with Delta Lake
- Enabling high-performance query engines, such as the Delta Engine, that accelerate SQL queries on massive datasets
- Optimizing execution for advanced analytics and ML with the DataFrame API

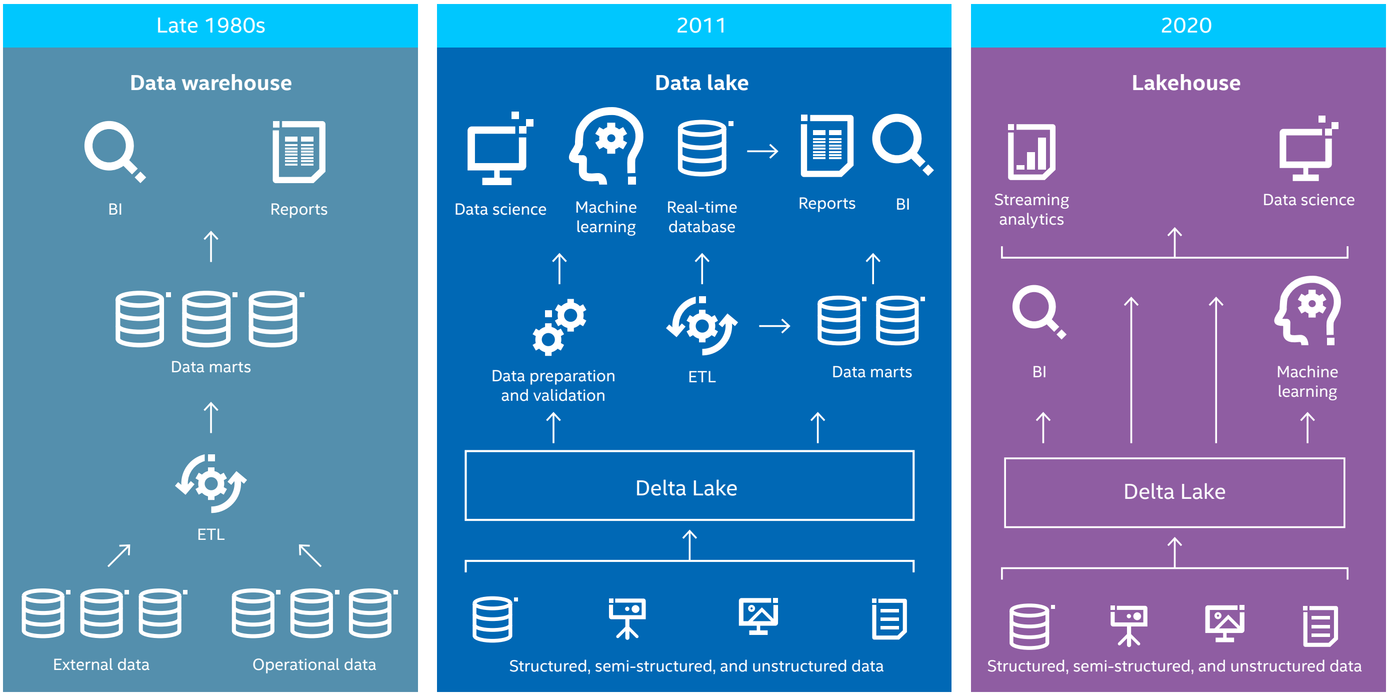

Figure 1. The progression from data warehouses and data lakes to Lakehouse architecture[1]

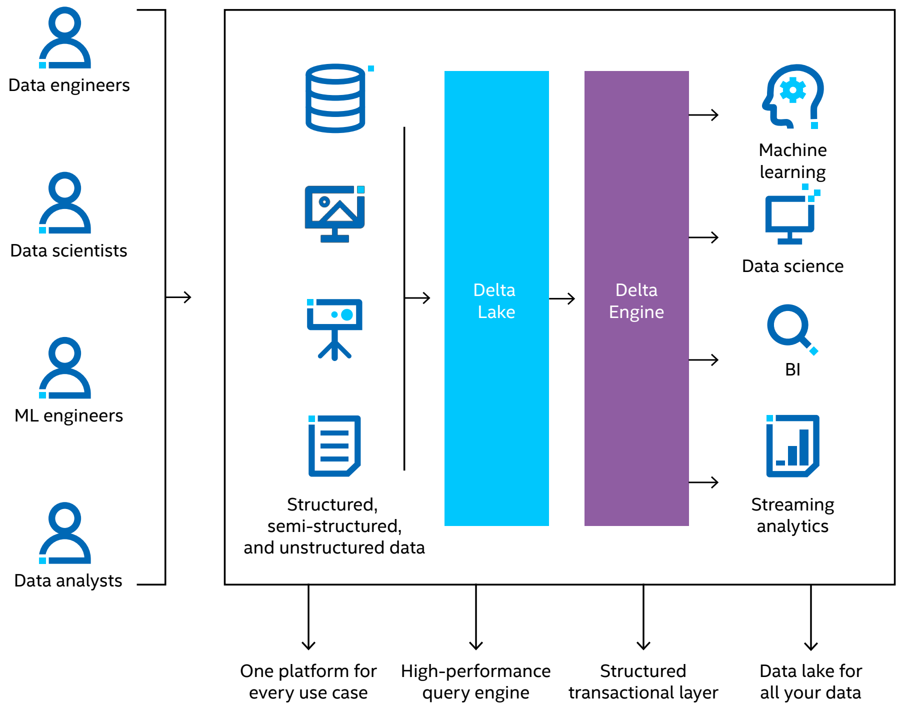
Data engineers
Data scientists
ML engineers
Data analysts
Structured, semi-structured, and unstructured data
Delta Lake
Delta Engine
Machine learning
Data science
BI
Streaming analytics
One platform for every use case
High-performance query engine
Structured transactional layer
Data lake for all your data

Figure 2. Intel Xeon Scalable processors provide performance benefits across data science and engineering workloads

Intel cloud infrastructure provides flexible performance options

As enterprises move workloads to cloud providers such as Azure, architects and software developers will want to be aware of the underlying hardware powering their cloud instances. Intel works across the industry to help ensure that software workloads are taking advantage of the additional performance from the latest, optimized Intel Xeon Platinum processors.

Intel has collaborated with Databricks to identify opportunities to extract additional performance for data science and engineering workloads when running on the latest Intel hardware available on the Azure platform.[2] Enterprises can realize these performance gains by choosing 2nd Generation Intel Xeon Platinum processor–based Azure VMs when configuring Azure Databricks clusters. By choosing the right instances for a given workload, these performance gains can translate into lower application TCO.

Enterprises can also realize performance gains by enabling Photon, a vectorized query engine that Databricks uses to provide fast query performance for SQL workloads. Photon is natively developed to use Intel Advanced Vector Extensions 2 (Intel AVX2). This native development provides data- and instruction-level parallelism on Intel processors to enhance SQL performance. By enabling Photon on Intel Xeon Platinum processor–based Azure VMs, enterprises can dramatically increase SQL query speeds.

Intel and Microsoft have worked together to bring enterprises a wide range of optimized VMs that run on 2nd Generation Intel Xeon Platinum processors. These VMs can handle virtually any size of workload. Table 1 lists Azure VMs that run on 2nd Generation Intel Xeon Platinum 8272CL processors.

The general-purpose Azure VMs feature local solid state drive (SSD) storage. They're ideal for workloads that benefit from low-latency local storage. Memory-optimized Azure VMs feature large RAM configurations and fast local SSD storage, which are ideal for memory-intensive applications. The memory-optimized Azure VMs that are optimized for Delta cache performance can help accelerate Delta Engine data reads.

Table 1. Azure VMs that run on 2nd Generation Intel Xeon Platinum 8272CL processors

General-Purpose Azure VMs	Memory-Optimized Azure VMs	Memory-Optimized Azure VMs for Delta Cache Acceleration
Standard_D4ds_v4	Standard_E20ds_v4	Standard_E8d_v4
Standard_D8ds_v4	Standard_E32ds_v4	Standard_E4ds_v4
Standard_D16ds_v4	Standard_E48ds_v4	Standard_E8ds_v4
Standard_D32ds_v4	Standard_E64ds_v4	Standard_E16ds_v4
Standard_D48ds_v4	Standard_E80ids_v4	
Standard_D64ds_v4	Standard_E4d_v4	
Standard_D4d_v4	Standard_E16d_v4	
Standard_D8d_v4	Standard_E20d_v4	
Standard_D16d_v4	Standard_E32d_v4	
Standard_D32d_v4	Standard_E48d_v4	
Standard_D48d_v4	Standard_E64d_v4	
Standard_D64d_v4		

Architects can use the Microsoft Azure Pricing Calculator to determine the best price-versus-performance options for their specific workloads.

Comparing Azure Databricks performance on multi-generation Intel Xeon Platinum processors

Intel recently created an environment in Microsoft Azure to test Azure Databricks data pipelining performance. The tests compared performance of VMs running on 1st Generation Intel Xeon Platinum 8171M processors to those running on 2nd Generation Intel Xeon Platinum 8272CL processors.

The testing environment used a workload that simulates a general-purpose decision support system by measuring query response, query throughput, and data-maintenance performance.

The first series of tests measured the performance of a cluster with 20 worker nodes or instances. The configuration was as follows:

- Databricks Runtime 9.0, which included Apache Spark 3.1.2, running on Ubuntu 20.04.1.
- The cluster consisted of 20 instances of Standard_E8s_v3 Azure VMs, each with 8 vCPUs and 64 GB of RAM, running in the East US 2 region.
- The underlying processors were 1st Generation Intel Xeon Platinum 8171M processors.
- The cluster included a single driver node that used the same configuration as the worker nodes.
- The tests were run against 1 TB and 10 TB datasets, and the total runtime was measured.

The tests were then repeated on a 20-node cluster running on 2nd Generation Intel Xeon Platinum 8272CL processors. The cluster configuration was identical except that each cluster node used a Standard_E8ds_v4 Azure VM type, with 8 vCPUs and 64 GB of RAM.

Intel also calculated the cost savings when running Databricks on the 2nd Generation Intel Xeon Platinum 8272CL processors. Table 2 lists the hourly price for each master node and worker node. It includes both the Standard_E8s_v3 Azure VMs (1st Generation Intel Xeon Platinum 8171M processors) and the Standard_E8ds_v4 Azure VM (2nd Generation Intel Xeon Platinum 8272CL processors). Intel used the Standard_E8s_v3 cost and performance as the baseline for both the 1 TB and 10 TB cluster tests.

For more information about determining worker instance type and size, see https://docs.databricks.com/clusters/cluster-config-best-practices.html#cluster-sizing-considerations.

Table 2. Hourly cost for Standard_E8s_v3 and Standard_E8ds_v4 Azure VMs

Azure VM	Hourly cost
Standard_E8s_v3 1st Generation Intel Xeon Platinum 8171M processors	$0.532
Standard_E8ds_v4 2nd Generation Intel Xeon Platinum 8272CL processors	$0.576

Performance and cost savings results

Figures 3 and 4 summarize the decision support workload performance test results. The results show that the 2nd Generation Intel Xeon Platinum 8272CL processor–based 20-node cluster outperformed the 20-node 1st Generation Intel Xeon Platinum 8171M processor–based cluster for both the 1 TB and 10 TB datasets with Photon enabled.

- The processing time for the 1 TB dataset was 55 percent faster.
- The processing time for the 10 TB dataset was 51 percent faster.

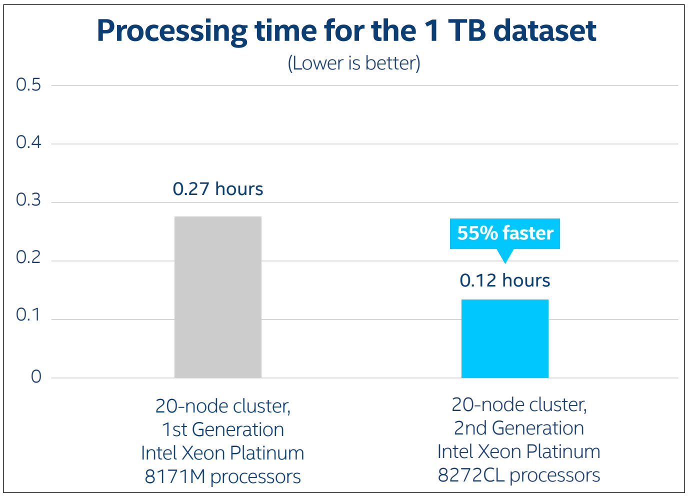

Figure 3. Decision support workload performance test results for the 1 TB dataset

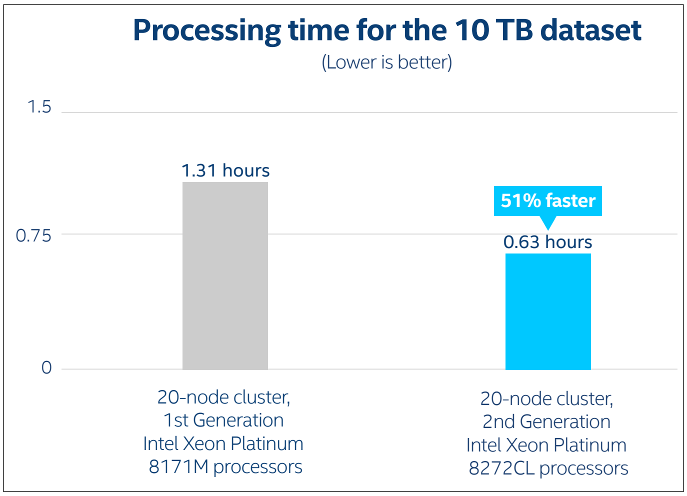

Figure 4. Decision support workload performance test results for the 10 TB dataset

Figure 5 summarizes the price/performance improvements. As with the decision support workload performance test results, the 2nd Generation Intel Xeon Platinum 8272CL processor–based 20-node cluster was more cost-effective than the 20-node 1st Generation Intel Xeon Platinum 8171M processor–based cluster. This cost savings applies for both the 1 TB and 10 TB datasets with Photon enabled:

- For the 1 TB cluster, the 2nd Generation Intel Xeon Platinum 8272CL processor–based cluster delivered a 55 percent overall savings.
- For the 10 TB cluster, the 2nd Generation Intel Xeon Platinum 8272CL processor–based cluster delivered a 51 percent overall savings.

Enterprises can use this information to help with sizing and cost comparisons to determine the best TCO for their use cases.

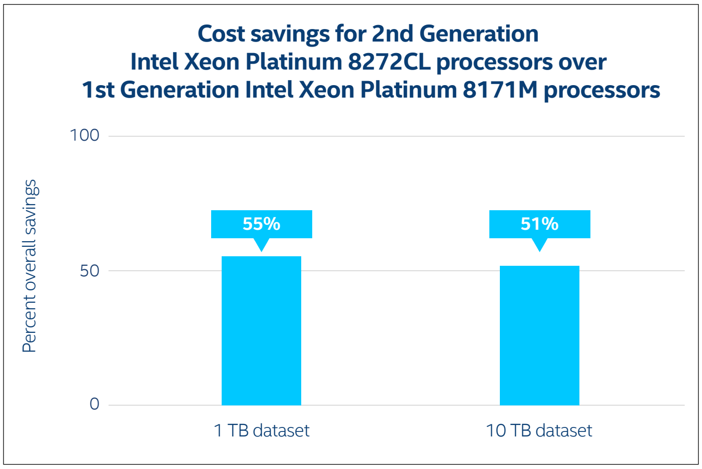

Figure 5. Cost savings for the 1 TB and 10 TB datasets running on 2nd Generation Intel Xeon Platinum 8272CL processor–based 20-node clusters

Discover the benefits of efficient data pipelining with Intel and Azure Databricks

Azure Databricks provides a modern alternative to isolated data warehouses and data lakes. With its unifying Lakehouse architecture running on Azure, Azure Databricks offers the benefits of both data warehouse and data lake technologies to provide the foundation for high-performance ML and AI workloads. Architects and software developers should be aware of the underlying hardware running their Azure Databricks workloads, as VMs based on Intel Xeon Platinum processors can provide a range of performance and cost benefits for any size of workload.

For more information, visit databricks.com.

Learn more

Discover Intel-optimized VM benefits in Azure: intel.com/microsoftazure

Discover the benefits of running workloads on Intel Xeon Scalable processors: intel.com/content/www/us/en/products/details/processors/xeon/scalable.html

Read more about the processors powering Azure VMs: https://azure.microsoft.com/en-us/pricing/details/virtual-machines/series/

Find out how to run Apache Spark jobs on an Azure Databricks workspace: https://docs.microsoft.com/en-us/azure/databricks/scenarios/quickstart-create-databricks-workspace-portal

Databricks Lakehouse frequently asked questions (FAQ): https://databricks.com/blog/2021/08/30/frequently-asked-questions-about-the-data-lakehouse.html

[1] Databricks. "What is a Lakehouse?" January 2020. https://databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html.

[2] Intel looked at the performance benefits for 20-node configurations. The Photon execution engine was enabled, which utilizes Intel Advanced Vector Extensions 2 (Intel AVX2). Pricing details are from https://azure.microsoft.com/en-us/pricing/calculator/ and https://azure.microsoft.com/en-us/pricing/details/databricks/. Pricing as of October 3, 2021. **Configurations: Databricks SQL configs**: spark. databricks. passthrough.enabled true, spark.databricks.adaptive.autoOptimizeShuffle.enabled true, spark.databricks.io.cache. maxMetaDataCache 10g, spark.databricks.io.cache. maxDiskUsage 100g, spark.databricks.delta.preview.enabled true. **System configs**: operating system: 20.04.1-Ubuntu, kernel: 5.4.0-1056-azure, Databricks 9.0 (includes Apache Spark 3.1.2, Scala 2.12). Tested by Intel October 3, 2021. **Baseline**: Azure-US-East-2, Standard_E8s_v3, 8 vCPUs, 20 instances, Intel Xeon Platinum 8171M processors at 2.60 GHz, 128 GB RAM/ instance, 256 GB storage/ instance (network or direct-attached), network bandwidth/instance: 32,000 (IOPS)/256 (MBps)/400 (cache), 8,000 storage bandwidth/instance. **New**: Azure-US-East-2, Standard_E8ds_v4, 8 vCPUs, 20 instances, Intel Xeon Platinum 8272CL processors at 2.60 GHz, 128 GB RAM/instance, 600 GB storage/instance (network or direct-attached), network bandwidth/instance: 154,000 (IOPS)/968 (MBps)/400 (cache), 8,000 storage bandwidth/instance.

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.

Printed in USA 1121/SMR/PRW/PDF Please Recycle 348490-001US