A Principled Technologies report: Hands-on testing. Real-world results.

up to **1.55x the throughput** for Bayes classification

up to **1.52x the throughput** for k-means clustering

Analyze more data per second on Apache Spark clusters with new Microsoft Azure VMs featuring 2nd Generation Intel Xeon Scalable Processors – Cascade Lake

Newer Dds_v4 series VMs featuring Intel Xeon Scalable 8259CL Cascade Lake processors enabled us to analyze data at a higher rate compared to older Intel Xeon E5-2686 v4 Broadwell processor-powered Ds_v3 VMs

Your company may use machine learning applications to develop new solutions and glean insights from the sorting, training, and analysis of massive amounts of data—but your competition does, too.

At Principled Technologies, we tested small, medium, and large[*] VMs of two Microsoft Azure VM series: the newer Dds_v4 series, featuring 2nd Generation Intel® Xeon® Scalable processors, also known as Cascade Lake processors, and older Ds_v3 series VMs, featuring Intel Xeon E5-2686 v4 processors, also known as Broadwell processors.

We created an Apache Spark™ cluster with these VMs and tested them with two machine learning workloads from the HiBench benchmark suite. Results from both benchmarking tests show that the newer Dds_v4 VMs processed data at a higher rate than the older VMs. This held true at each instance size we tested.

With new Cascade Lake processor-powered Dds_v4 VMs running your machine learning workloads, your company could develop solutions and insights sooner than if you chose older Ds_v3 VMs—and maybe even faster than your rivals down the street.

[*] *Apache documentation refers to this as the Master node. We call it the Primary node.*

How we tested

We tested two series of memory-optimized VMs for Microsoft Azure:

- Newer Dds_v4 series VMs featuring Cascade Lake processors
- Older Ds_v3 series VMs featuring Broadwell processors
  - Note: Ds_v3 VMs come in multiple processor configurations. For our testing, we chose these processors only.

We compared these VMs across three sizes: small VMs (8 vCPUs with two 4-core executors), medium VMs (16 vCPUs with four 4-core executors), and large VMs (64 vCPUs with sixteen 4-core executors). We chose these sizes to represent a spectrum of workloads companies may need. We increased the number of workload executors to scale the amount of work relative to the available hardware and tuned the workloads to utilize as much CPU as possible.

For each environment, we used five VMs to create a 1+4 cluster of identical VMs with one VM acting as a Primary node* and the other four VMs acting as the Worker nodes. We used Hadoop Distributed File System as the cluster storage file system. Figure 1 proves a visual breakdown of the VMs we tested.

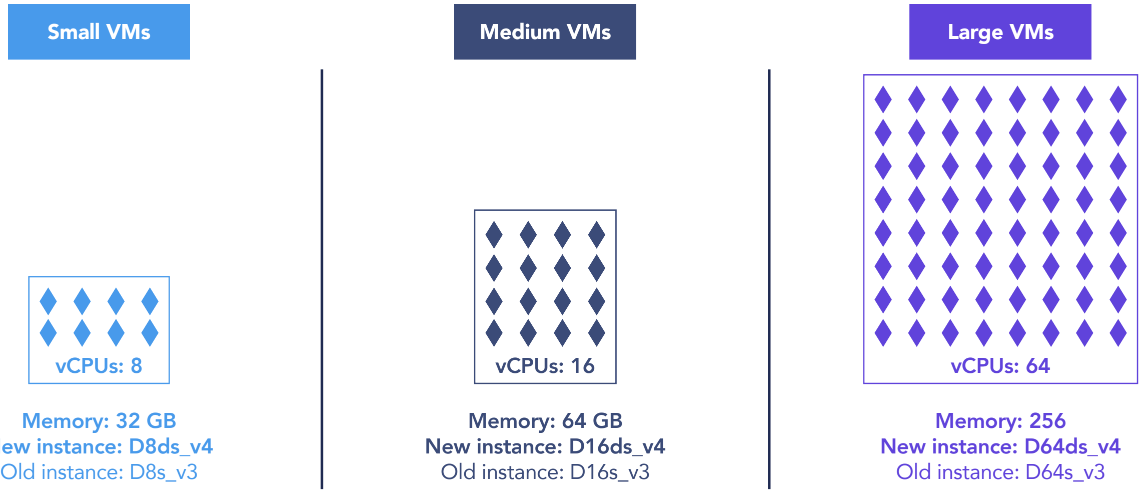

Figure 1: Specifications for the Microsoft Azure instances we used in testing. Note that we tested each instance in the East US (Zone 1) region Source: Principled Technologies.

Benchmarking

Your company's machine learning workflow likely involves processing and analyzing unstructured data from disparate sources. We used two tests from the HiBench benchmark suite to assess each instance series' large-scale machine learning performance:

- **k-means clustering:** The HiBench implementation of the well-known insight-discovery and data-mining algorithm
- **Naive Bayesian classification:** The training portion of the Naive Bayesian classification, which is a popular algorithm for knowledge discovery and data-mining

Because each of these benchmarks represented a different method for manipulating data, we were able to view multiple perspectives on each cluster's large-scale machine learning performance.

**Apache documentation refers to this as the Master node. We call it the Primary node.*

Our results

Small VMs

Small businesses with large-scale machine learning needs may find that new Dds_v4 VMs for Microsoft Azure can satisfy their current level of work while leaving room to grow in the future. As Figure 2 shows, for both the k-means clustering and Naive Bayesian classification workloads, the Cascade Lake processor-powered D8ds_v4 VM cluster processed data at a higher rate than the Broadwell processor-powered D8s_v3 VM cluster.

Small VMs comparison: relative throughput

Higher is better

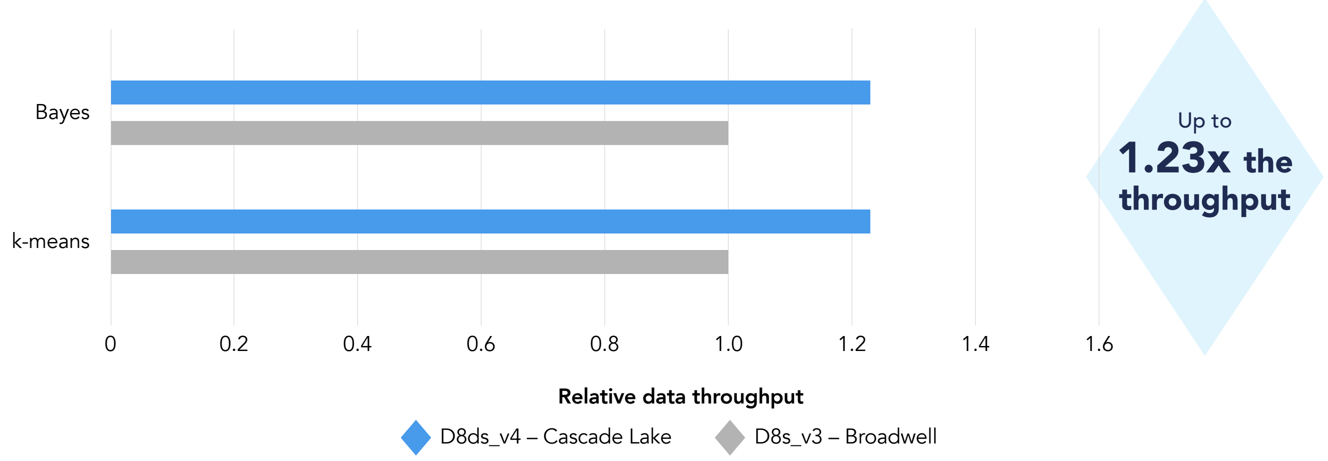

Figure 2: Normalized comparison of the average data throughput each small VM achieved in the Naïve Bayesian classification and k-means clustering workloads. Higher throughput is better. Source: Principled Technologies.

Medium VMs

Mid-sized business could find more than middle-of-the-road big data performance with 16-vCPU Dds_v4 VMs. Figure 3 shows that the D16ds_v4 VM cluster processed 1.31x the rate of data as the older D16s_v3 VM cluster on the k-means clustering workload, and 1.46x the rate of data on the Naive Bayesian classification benchmark.

Medium VMs comparison: relative throughput

Higher is better

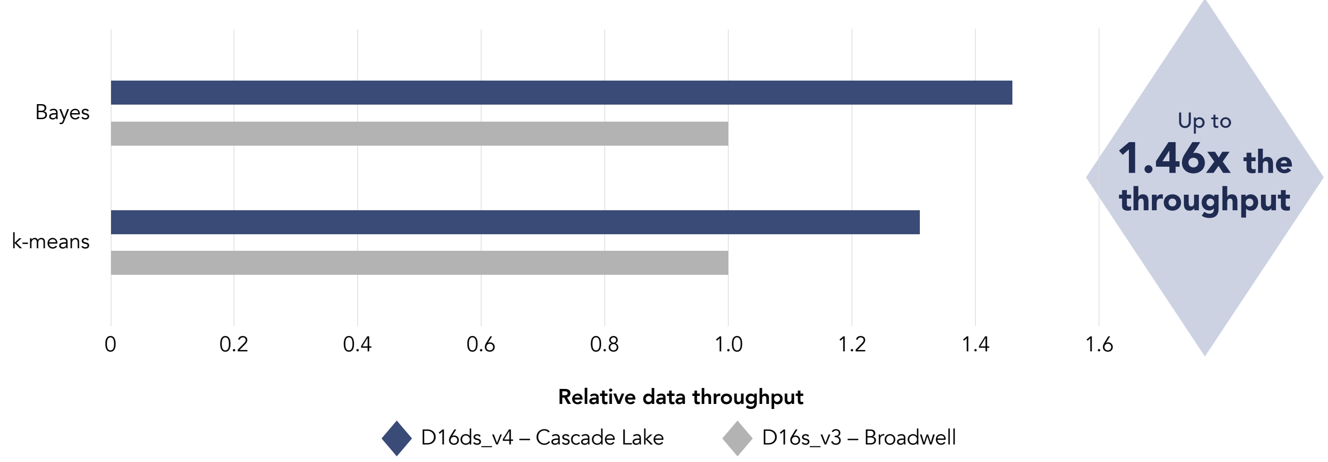

Figure 3: Normalized comparison of the average data throughput each medium VM cluster achieved in the Naïve Bayesian classification and k-means clustering workloads. Higher throughput is better. Source: Principled Technologies.

Dds_v4 series VMs for Microsoft Azure

According to Microsoft, Dds_v4 series VMs offer several advantages that may contribute to better performance than the Ds_v3 series:[1]

- Guaranteed 2nd Generation Intel Xeon Scalable processors for every VM
- 50 percent larger default disk drives
- Higher IOPS on default disk drives

To learn more, visit https://docs.microsoft.com/en-us/azure/virtual-machines/ddv4-ddsv4-series.

Large VMs

Large businesses that gather and process many terabytes of data could sort through and analyze more data in a given session with Dds_v4 VMs for Microsoft Azure. Figure 4 shows that the Cascade Lake processor-powered D64ds_v4 VM cluster processed data at 1.52x the rate of the Broadwell processor-powered D64s_v3 VM cluster during the k-means clustering algorithm test, and 1.55x the rate during the Naïve Bayesian classification test.

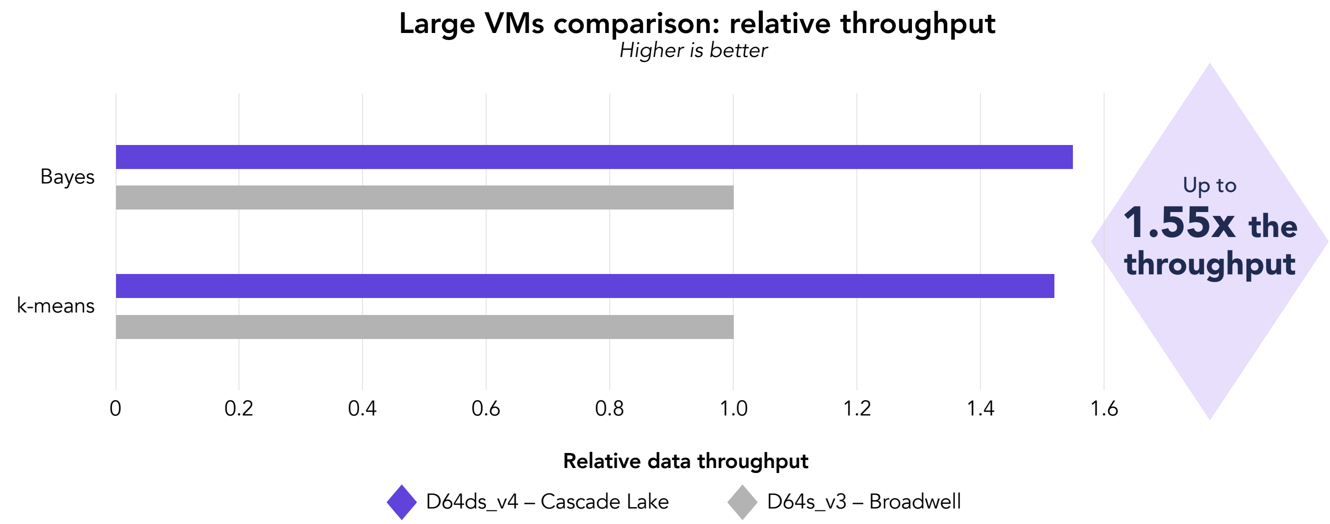

Figure 4: Normalized comparison of the average data throughput each large VM cluster achieved in the Naive Bayesian classification and k-means clustering workloads. Higher throughput is better. Source: Principled Technologies.

Better performance and better value

In our tests, the newer Dds_v4 series VMs featuring Cascade Lake processors processed data at 1.23 to 1.55 times the rate of the older Ds_v3 series VMs featuring Broadwell processors. Because the Dds_v4 VMs cost just 1.17 times the price of the Ds_v3 VMs, they present a better investment value for your Apache Spark big data workloads.[2]

Better performance with larger VMs

In our tests, the margin by which newer Dds_v4 VMs outperformed older Ds_v3 VMs increased significantly with larger VMs. While the small Dds_v4 VMs processed up to 1.23x the data of the small Ds_v3 VMs, the medium Dds_v4 VMs processed up to 1.46x the data as their counterparts. The large Dds_v4 VMs processed even more data at up to 1.55x the rate of the Ds_v3 VMs.

1.23x the throughput *for small VMs*

1.46x the throughput *for medium VMs*

1.55x the throughput *for large VMs*

Conclusion

If machine learning is key to your organization's mission, you're more likely to thrive if your Apache Spark VM clusters can process data at a higher rate than the competition.

In our hands-on tests, newer Microsoft Azure Dds_v4 series VM clusters featuring Cascade Lake processors handled data at a higher rate than older Ds_v3 VM clusters featuring Broadwell processors:

- k-means clustering algorithm: Up to 1.52x the rate
- Naïve Bayesian classification: Up to 1.55x the rate

With VM clusters powered by newer processors, your organization could sooner gain insights and develop solutions from the data you collect, and potentially ready itself for growth.

[1] "Ddv4 and Ddsv4-series," accessed December 4, 2020, https://docs.microsoft.com/en-us/azure/virtual-machines/ddv4-ddsv4-series.

[2] "Windows Virtual Machine Pricing," accessed December 4, 2020, https://azure.microsoft.com/en-us/pricing/details/virtual-machines/windows/.

Read the science behind this report at http://facts.pt/mqvkSpH ▶

Principled Technologies®

Facts matter.®

This project was commissioned by Intel.