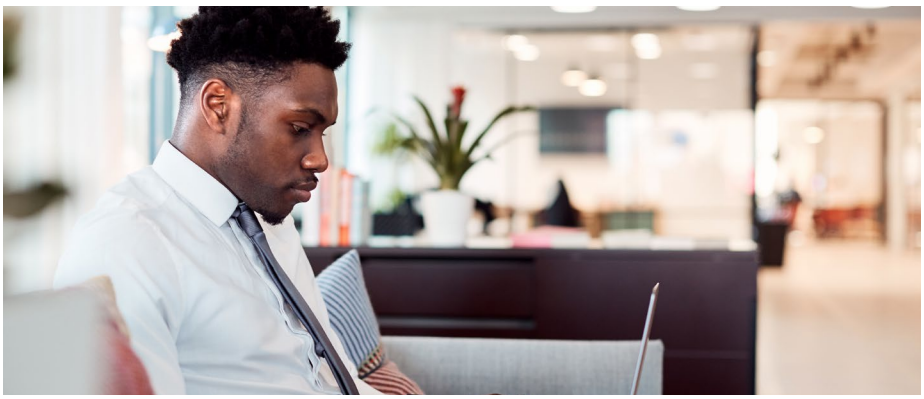
Save on Time and Cost by Choosing Google Cloud Platform N2 VMs with 2nd Gen Intel® Xeon® Scalable Processors and Databricks Photon Query Engine

Use Photon to Maximize Decision Support Database Performance on n2-highmem-8 VMs featuring Intel Xeon Scalable Processors

For organizations that store, access, and analyze vast amounts of structured and unstructured data, the Lakehouse Platform from Databricks provides a unique combination of data warehouse and data lake features. The platform also includes Photon, a vectorized query engine that is designed to speed SQL query performance. According to a summary from Databricks, Photon benefits include:

- "Supports SQL and equivalent DataFrame operations against Delta and Parquet tables.
- Expected to accelerate queries that process a significant amount of data (100GB+) and include aggregations and joins.
- Faster performance when data is accessed repeatedly from the Delta cache.
- More robust scan performance on tables with many columns and many small files.
- Faster Delta and Parquet writing using UPDATE, DELETE, MERGE INTO, INSERT, and CREATE TABLE AS SELECT, especially for wide tables (hundreds to thousands of columns).
- Replaces sort-merge joins with hash-joins."[1]

Speedier queries translate to faster time to business insights and less VM uptime to pay for. To test Photon on Google Cloud Platform (GCP) N2 VMs, we used a decision support benchmark, which measured data warehousing performance by running a set number of queries and recording the time to complete them. When we compared the performance of Photon-enabled n2-highmem-8 VMs featuring 2nd Gen Intel Xeon Scalable processors to that of same VMs without Photon, we found that the Photon-enabled N2 VMs completed queries in less time on 1TB and 10TB datasets—all while reducing cost in both scenarios.



Databricks

N2 VMs with Photon enabled completed decision support database queries up to 3.6 times as fast

as N2 instances without Photon

Running decision support databases on N2 VM instances without Photon cost up to 2.3 times as much

as N2 VMs with Photon

Speed Time to Insight with Photon

To determine how Photon can enhance query performance, we tested eight-vCPU n2-highmem-8 VMs with and without Photon. Figure 1 shows how the N2 VM cluster with Photon completed a 1TB dataset 3.3 times as fast as the same cluster without Photon, and completed a 10TB dataset 3.6 times as fast.

Normalized Processing Time to Complete Queries on Databricks With vs. Without Photon

Relative time to complete | Lower is better

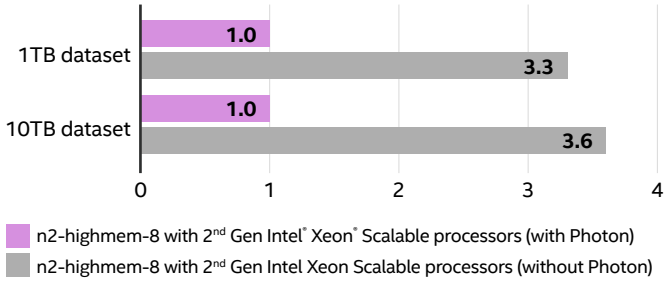


Figure 1. The relative processing time to complete decision support benchmark queries with Photon compared to without Photon on GCP n2-highmem-8 VMs on 1TB and 10TB datasets.

Enable Photon for a Better Value

While performance improvements sometimes come at a higher price, we found that the faster processing times with Photon translate to less VM uptime costs. Figure 2 shows that compared to the N2 cluster with Photon, the cluster without Photon costs 2.1 times more when analyzing a 1TB dataset and 2.3 times more when analyzing a 10TB dataset.

Normalized Cost to Run a Decision Support Workload With vs. Without Photon

Relative cost | Lower is better

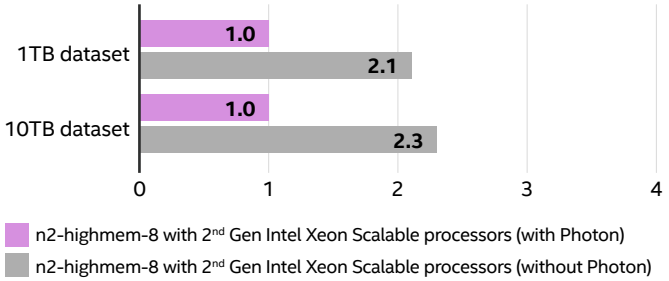


Figure 2. Normalized price/performance to run a decision support workload against a Databricks environment on GCP n2-highmem-8 VMs on both 1TB and 10TB datasets.

Conclusion

If your organization supports decision making databases with Databricks, the Photon query engine on GCP n2-highmem-8 VMs can reduce query completion time and deliver a better value. With Photon, these eight-vCPU VMs completed a decision support database workload up to 3.6 times as fast as those without Photon. These performance improvements led to a better value, with N2 VMs without Photon costing up to 2.3 times as much as their Photon-enabled counterparts. For speedier performance and cost savings, choose GCP N2 VMs featuring 2nd Gen Intel Xeon Scalable processors with Photon enabled.

Learn More

To begin running your Databricks clusters with Photon enabled on GCP N2 VMs with 2nd Gen Intel Xeon Scalable processors, visit https://cloud.google.com/compute/docs/general-purpose-machines.

1. Databricks, "Photon," accessed April 12, 2022, https://docs.databricks.com/runtime/photon.html.

Tests by Intel in March 2021 on GCP us-central1 (Iowa). All configurations: 21 instances (20 workers + 1 master), N2-highmem-8 instances with Intel Cascade Lake CPUs, 08 vCPUs, 128GB RAM, 25 Gbps, 500GB remote SSD+0.75TB local SSD, 240-1200/240-1200 (R/W remote SSD) 9360/4680 (R/W local SSD) Ubuntu 20.04.3 LTS kernel 5.4.170+, Databricks 10.3. Spark config: 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. Total cluster cost per run as of Mar 2022: w/Photon 1TB: $6.44; w/Photon 10TB: $33.11, w/o Photon 1TB: $13.95; w/o Photon 10TB: $78.10.

Printed in USA 0522/JO/PT/PDF US002 ♻ Please Recycle