Run compute-intensive Apache Hadoop big data workloads faster with Dell EMC PowerEdge R640 servers

A Hadoop cluster of PowerEdge R640 servers powered by 2nd Generation Intel Xeon Scalable processors completed three compute-heavy big data workloads in less time than previous-generation Dell EMC PowerEdge R630 servers by processing more data per second

Analyze more data per second

Up to 112% greater throughput while analyzing words in a large document

Identify topics in documents more quickly

Up to 52% less waiting for document analysis

Your organization likely generates large volumes of data from numerous sources continuously. This data can range from how long users are on a web page to the length of routine sales team video calls. Extracting insight from this disparate information often requires running complex, compute-intensive workloads quickly on multiple data sets.

Aging servers typically cannot deliver the speed that newer servers can offer for these compute-intensive workloads. Current-generation servers can deliver a performance improvement that helps your organization now and allows you to continue accumulating and using data effectively. Faster servers can process and analyze data more quickly, so marketing teams, for example, can more quickly determine who to target for their next email campaign.

In our data center, an Apache Spark™ cluster of current-generation Dell EMC™ PowerEdge™ R640 servers featuring 2nd Generation Intel® Xeon® Scalable processors outperformed a cluster of previous-generation Dell EMC PowerEdge R630 servers in three compute-intensive, big data workloads. The workloads identified topics in a large document, classified information to make a prediction, and counted words in a data set. Moving these workloads to new PowerEdge R640 servers and getting better performance can help your organization can meet today's demands and offer the computing power necessary to face the challenges of tomorrow.

In addition to newer, faster Intel processors to run queries and algorithms, the PowerEdge R640 servers had more drive bays than their predecessors. More drive bays allow you to add more storage to each server and store more data, which could help prevent server bottlenecks and promote speedy access to databases.

About the Dell EMC PowerEdge R640

The Dell EMC PowerEdge R640 is a dense 1U, two-socket server. It features 24 DDR4 DIMM slots and up to 76.8 TB of storage.

To learn more about the Dell EMC PowerEdge R640, visit https://www.dell.com/en-us/work/shop/povw/poweredge-r640.

About 2nd Generation Intel Xeon Scalable processors

The latest from Intel, the 2nd Generation Intel Xeon Scalable processor platform features a wide range of processor types, including Bronze, Silver, Gold, and Platinum, to support the workloads you run. According to Intel, the 2nd Generation Intel Xeon Scalable platform can handle a variety of workloads, including enterprise, cloud, HPC, storage, and communications.[1] This new processor line also supports a new memory and storage technology to further accelerate workloads, Intel Optane™ DC persistent memory.

To learn more about the 2nd Generation Intel Xeon Scalable processor family, visit https://www.intel.com/content/www/us/en/products/docs/processors/xeon/2nd-gen-xeon-scalable-processors-brief.html.

Why it helps to regularly replace servers in your data center

Efficient use of data can help organizations make more informed choices in many areas: operations, customer service, and supply chain, among others. But when organizations don't have the tools necessary to analyze the data they're generating, they run the risk of missing competitive advantages and wasting resources to store the data. Moving big data workloads to current-generation servers can help you meet the challenges of expanding data sets. Newer servers typically introduce improved or new technology, such as greater storage capacity, increased network speeds, and faster processors.

In addition, newer servers often have new or improved management features that make it easier for IT staff to roll out firmware updates, monitor the health of physical and virtual layers, and set up new applications and software.

Replacing servers can also help IT teams holistically, allowing them serve as the strategists for organizational transformation rather than as merely a support team. By moving compute-heavy, big data workloads to newer servers, IT teams can focus on building forward-looking models for IT service delivery, application modernization, and other key data center initiatives.

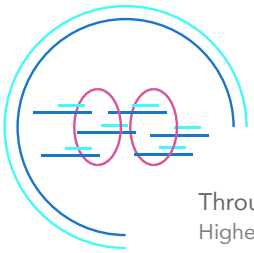
Run compute-heavy Hadoop big data workloads more quickly

Big data workloads running on Hadoop framework can generate valuable insights that help organizations predict behaviors or outcomes. Decision-makers and their teams can use the insight to orchestrate transformative initiatives, such as targeted email campaigns that drive sales or customer feedback analysis that improves product quality. Current-generation PowerEdge R640 servers could allow business units and data analysts to work with large data sets more quickly than those in organizations that continue to run big data analysis workloads on previous-generation servers.

To see how the two solutions could handle real-world big data work, we ran three HiBench big data tests on them:

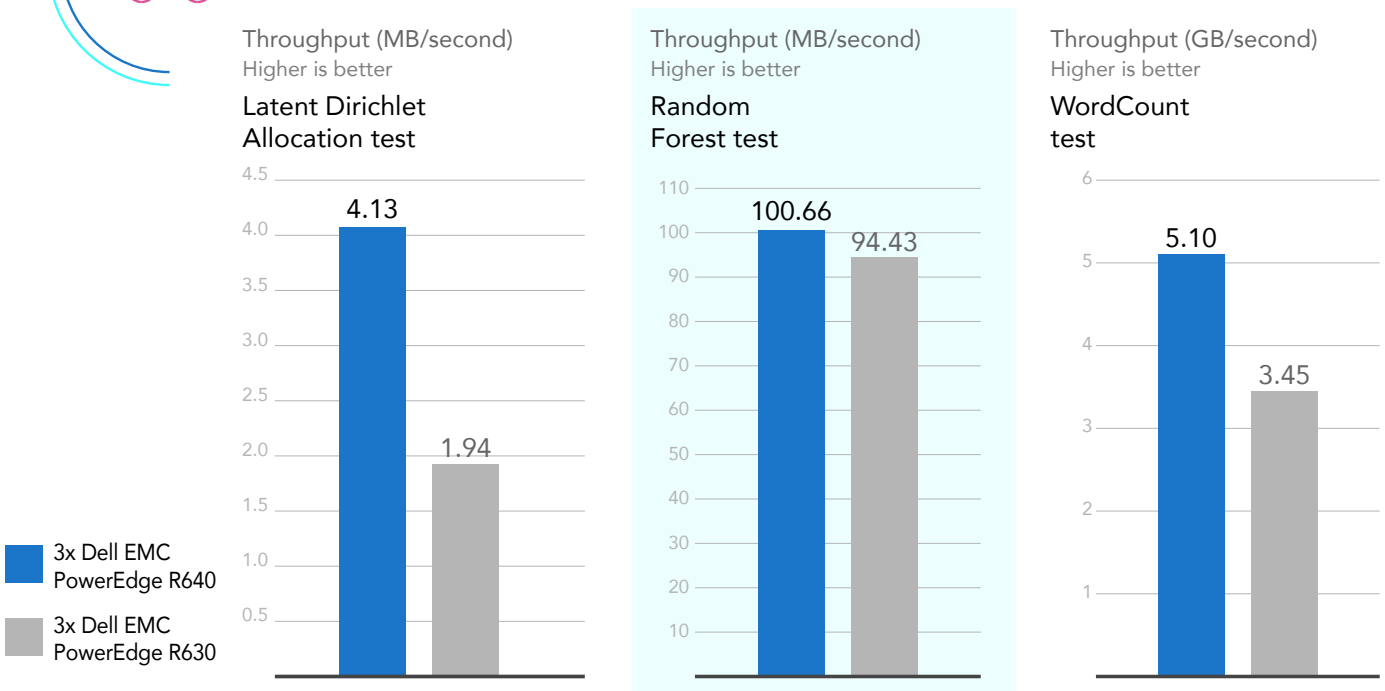
- **Latent Dirichlet allocation (LDA)**: LDA is a technique to dynamically identify the topics discussed in a given document. It analyzes the words in a document, temporarily categorizes them, refines the topic of the document with the previously identified categories, and summarizes the document. A business could use LDA, for example, to organize customer reviews on a product or service.
- **RandomForest (RF)**: Organizations can use random forests to make predictions by running multiple decision trees with slightly different weights and comparing the outcome between the trees to prevent overfitting—this can greatly increase the accuracy of a decision tree in classification and regression. For example, a bank or financial institution could use RF to make credit risk predictions.
- **WordCount**: This workload tallies the occurrence of each word in the data. The benchmark generates the random text from RandomTextWriter, which is a program that uses map/reduce to just run a distributed job with no interaction between tasks, and each task writes a large, unsorted random sequence of words. WordCount is "representative of another typical class of real-world MapReduce jobs - extracting a small amount of interesting data from large data set."[2]

The cluster of current-generation Dell EMC PowerEdge R640 servers powered by 2nd Generation Intel Xeon Scalable processors completed the three workloads more quickly than the previous-generation solution. The current-generation PowerEdge R640 servers ran the LDA workload with a throughput of more than 4 MB per second—more than double the throughput of the previous-generation solution. Processing more data per second could allow more of your business units to access and use the data. The chart below shows the throughput for both solutions in each test.



Analyze more data per second

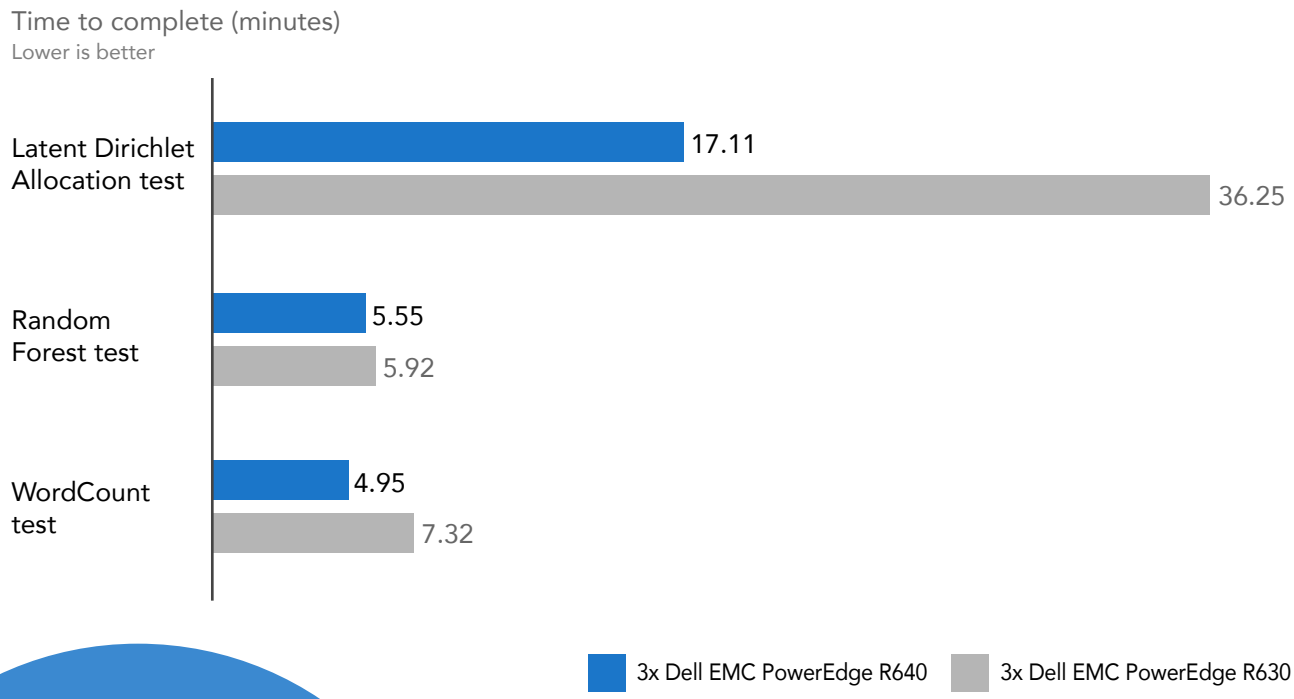
Up to 112% greater throughput while analyzing words in a large document



The current-generation Dell EMC PowerEdge R640 solution powered by 2nd Generation Intel Xeon Scalable processors needed just over 17 minutes to process 4.5 GB of data for the LDA test. Compared to the previous-generation solution, which needed 36 minutes, the solution completed the workload in less than half the time. Not only could you deliver analysis to decision makers more quickly with the PowerEdge R640, but you could use the extra time, for example, to re-run the LDA workload to ensure accuracy. The chart below shows the times to complete all three tests for both solutions.

Identify topics in documents more quickly

Up to 52% less waiting for document analysis



About HiBench

Intel HiBench 7.1 is a big data benchmark suite for Apache Hadoop. Some tools in the suite are synthetic micro-benchmarks while others are real-world Hadoop applications. The output of the tools can demonstrate a solution's processing speed, throughput, bandwidth, CPU utilization, data access patterns, and other metrics as they relate to processing big data workloads.

For more information on HiBench, visit https://github.com/Intel-bigdata/HiBench.

Conclusion

Efficiently running compute-heavy, Apache Hadoop big data workloads today might not translate to continued quality performance for growing data sets. Moving compute-intensive, Hadoop big data workloads to current-generation Dell EMC PowerEdge R640 servers powered by 2nd Generation Intel Xeon Scalable processors could allow your organization to better meet the data analysis challenges of today and have the resources to support growth. In our data center, a Hadoop cluster of PowerEdge R640 servers completed three big data workloads in less time by delivering greater throughput than a cluster of previous-generation Dell EMC PowerEdge R630 servers. Faster analysis of large data sets means getting insight into your organization, products, and services sooner, which could help your organization grow and beat its competition.

---

1 Intel, "2nd Gen Intel Xeon Scalable Processors Brief," accessed November 7, 2019, https://www.intel.com/content/www/us/en/products/docs/processors/xeon/2nd-gen-xeon-scalable-processors-brief.html.

2 "Intel-bigdata/HiBench," accessed November 5, 2019, https://github.com/Intel-bigdata/HiBench.

Read the science behind this report at http://facts.pt/gv58wjr ►

Principled Technologies®

Facts matter.®

This project was commissioned by Dell EMC.