DELL EMC
PowerEdge

THE **COMPUTE** INFRASTRUCTURE NEEDED FOR PREDICTIVE **ANALYTICS**

Introduction

Modernized organizations are leveraging the transformative benefits of predictive analytics to propel their business forward. Predictive analytics arm you with knowledge and insights that can make the difference between profits and losses. They use modeling, machine learning, and artificial intelligence to analyze current data and make predictions about the future. Every industry can benefit from predictive analytics. Those who ignore the potential here are squandering a huge opportunity to grow or create better experiences.

The recent dramatic progress in AI-based predictive analytics and machine learning provides organizations with the power to unlock tremendous value from previously hard to decipher data. Through the power of AI-enabled infrastructure, you can harness the predictive analytics/machine learning wave and gain business advantages that will not only keep you in the game, but at the leading edge. Automated server capabilities and the right server for the job will get you where you need to go. This eBook focuses in on how server automation, along with the right servers for the workload, empowers companies to make great strides in the world of machine learning and predictive analytics.

Predictive Analytics and Machine Learning

A recent ESG research paper, Three Transformational Compute Technologies Verified to Accelerate AI and Business Value, segmented organizations according to their level of advancement in artificial intelligence capabilities. The three criteria used as benchmarks were server automation, accelerators, and converged/hyperconverged infrastructure. Having automated server capabilities helped push organizations into the Most Advanced (Stage 3) category. While only about half (55%) of Stage 1 companies are developing, deploying, and tuning AI models in production for predictive analytics, a whopping 74% of Stage 3 businesses are pursuing these goals.[1]

Why does this matter? There are more use cases for predictive analytics and machine learning than there is room in this eBook. Here are just a few examples:

- Creating recommendation engines for customers to keep them purchasing your products in a way that is relevant to them.
- Factories and supply chains can utilize analytics to predict when failures will occur, thus circumventing downtime.
- Predict the optimal price to sell a product at based on past data trends such as foot traffic and weather forecasts.

These scenarios do not even scratch the surface of what you can do with predictive analytics. By automating your servers, you free up employee time to focus on advancing AI initiatives rather than babysitting infrastructure. Your staff can then work on consolidating your information into data lakes They will use applications like Hadoop and Spark to solve problems and glean actionable predictions from that immense quantity of data you've been collecting for years. And they will run all of this on servers with excellent processing, storage, and memory power. When you make predictive analytics and modern infrastructure a priority, you will become more like a Stage 3 organization. Imagine if you could harness these benefits.

ESG AI Maturity Stages

Stage 1

(**42%** of organizations in the study): Low levels of automation, very limited use of accelerators, and/or little to no converged/HCI-based infrastructure for AI.

Stage 2

(**33%** of organizations): Moderate levels of automation, some use of accelerators, and/or some converged/HCI-based infrastructure for AI.

Stage 3

(**24%** of organizations): High levels of automation, broad use of accelerators, and/or high use of converged/HCI-based infrastructure for AI.

Stage 3 organizations:[2]

are **2.6x** more likely than Stage 1 organizations to lead the competition in business intelligence and analytics.

are **2x** more likely to see shorter time to value.

improve decision speed with AI by nearly **2x** more than Stage 1 organizations.

improve decision accuracy with AI by nearly **2.6x** more than Stage 1 organizations.

see **2.7x** the average cost reduction through automation of business processes and/or operations as compared to Stage 1 organizations.

Applications to Help You Win at Predictive Analytics: Spark Running on Hadoop

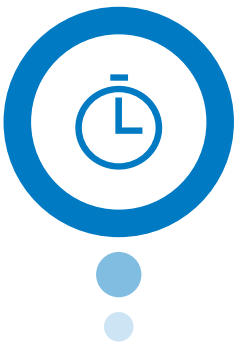
Hadoop is an open-source platform that will provide the base for your predictive analytics and machine learning activities. When running predictive analytics, your first challenge (after figuring out what business problems you want to solve or market opportunities you want to exploit) is understanding how to gather all your relevant data together. Gathering all the relevant data together enables you to pool your resources so that the predictions you generate are as accurate as possible. The biggest benefit of Hadoop is that it is a highly distributed file system that allows you to store any piece of data in native format without transformation, making it ideal not only for structured data, but also for unstructured data.

Some sample use cases for Hadoop:

- Retailers better serve their clientele by using it to help analyze structured and unstructured data to help them tailor their offerings more accurately.
- Hadoop serves data to the applications that allow financial services companies to prevent fraud.
- Logistics companies use Hadoop-powered analytics to assess whether preventive maintenance is needed on their equipment.

It's important to keep in mind that Hadoop must be partnered with another application to run true predictive analytics and machine learning. Often, that application is Spark. Spark is an open-source project with enhanced functionality. In Hadoop you have map and reduce functions. The map function takes a data set and converts it into another data set. This new set breaks down elements into key/value pairs. The reduce function then combines these key/value pairs into a smaller set of key/value pairs. In Spark you additionally can join, cluster, filter, and add or delete data. This offers you additional opportunities to manipulate your data. Spark offers high-speed analytics and has a machine learning library built in. Spark's in-memory processing engine ensures real-time predictive results. The important thing to remember is that Spark and Hadoop work in partnership to deliver these benefits to your organization.

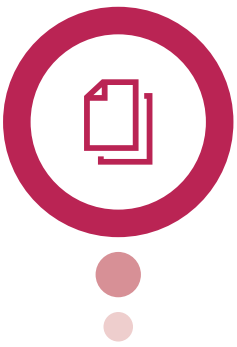
Important features of implementing Spark:



Real-time data querying

A large, supportive community

Rapid stream processing of low latency data; runs workloads 100x faster than Hadoop alone[3]



Rich libraries

Automated, Well-Rounded Servers Are Needed to Run Predictive Analytics

To get the most out of the Hadoop and Spark pairing, you'll need:

1. Automated servers that free up time and resources for machine learning pursuits and deliver insights more quickly to your customers.
2. Powerful and plentiful core processors for crunching all that data and running those algorithms.
3. Substantial storage to ensure that Hadoop runs without limitations.
4. Serious server memory to allow Spark's in-memory processing to function optimally.

Automate Your Servers

What does server automation have to do with running predictive analytics and machine learning? A lot, actually. When you replace your aging servers with modern servers with automation capabilities, you can expect:[4]

43% faster deployments

43% higher systems reliability

38% faster updates of applications

37% less time spent on routine, manual IT infrastructure management

This translates into quicker deployment of insights, and more time for staff to focus on delivering them. You want your analytics servers up and running lightning-fast, and you want them back online should they go down. You want to be able to deploy your Hadoop clusters with efficiency, and to update Hadoop and Spark as quickly as possible when the need arises. The more you automate, the less time you spend on manual maintenance processes that can slow down your analytics value journey.

Find the Right Well-Rounded Server for Hadoop and Spark

What you'll need for the job of predictive analytics using Hadoop and Spark is a server that has it all: strong showings in processing, storage, and memory.

Processors are the brains behind your operation, analyzing and interpreting all the instructions that other hardware and software throw their way. In machine learning, the processor is tasked with executing the logic in your algorithm. It carries a lot of weight; core count absolutely matters. It's important to make sure you choose infrastructure that can support these processing-intensive workloads.

> Spark recommends provisioning at least 8-16 cores per machine in order to crunch all that data needed for your predictive analytics.[5]

Hadoop is a storage-hungry application that does large-batch processing. It processes vast amounts of data ranging into the petabytes. To run Hadoop, make sure to provision servers that don't come up short on storage. Although Hadoop follows a distributed model, each individual server or node needs adequate storage potential on local disks for it to run smoothly. It's important to understand that there are multiple types of nodes in Hadoop. This eBook will focus on offerings for the edge node, which acts as the gatekeeper to the cluster, and worker nodes, which drive the actual processing of your data.

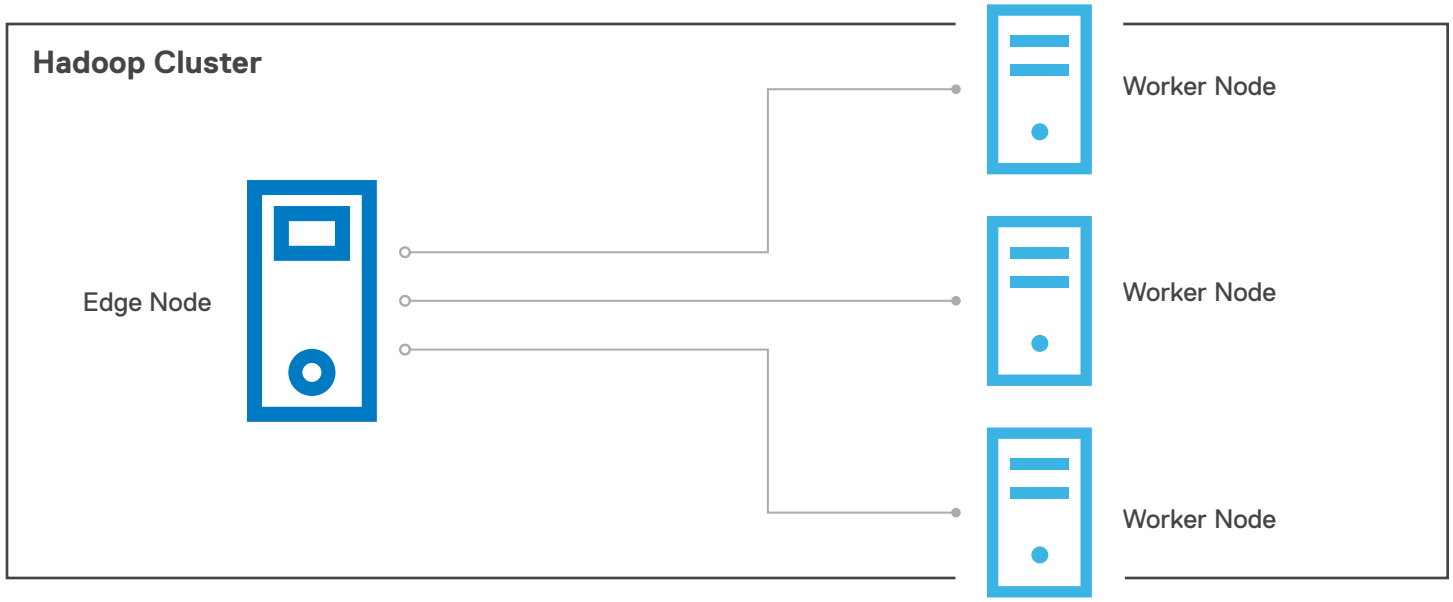
Spark, on the other hand, is memory intensive. Although Spark's minimum recommendation is 8 GB of memory per machine, a much larger amount of memory is ideal. Some experts recommend at least 128 or 256 GB. Spark recommends allocating at most 75% of available memory for Spark and leaving the rest for the operating system and buffer cache.[6] Memory is particularly key for this application because it caches data in memory across multiple parallel operations.

In summary, you need automated workhorse servers that can do it all when it comes to exploring predictive analytics and machine learning. Read on to see the Dell EMC servers built for these resource intensive workloads.

PowerEdge Servers for Your Predictive Analytics Needs

First, all product recommendations below are prefaced with the fact that they offer extremely compelling automation options. OpenManage, the Dell EMC premier server management platform, offers RESTful APIs. The APIs allow you to script many aspects of your server deployment, maintenance, and provisioning. You can write a script, set it, and forget it. The OpenManage portfolio also supports the Ansible framework, which offers modules allowing IT to automate in a multi-vendor scenario. These modules are delivered as pre-written scripts for fine-tuned automation.[7]

To build a Hadoop cluster you will need multiple node types. The diagram below will focus on a single edge node and multiple worker nodes. The edge node manages access to Hadoop cluster and acts as the entry point between your Hadoop cluster and the outside network. The edge node is not in charge of computationally intense work. That task falls to the worker nodes. Worker nodes take on the heavy lifting of processing the data for you and require generous memory and storage, as well as processing power. For detailed information on Hadoop, you can read more here.



Edge Node

PowerEdge R640

Although the edge node is not recruited for intense data crunching, it still needs to have powerful processors, storage, and memory to keep the cluster running smoothly. The Dell EMC PowerEdge R640 is just right for this job; neither too much nor not enough for this role. The R640 lets you scale compute resources with up to two 2nd Generation Intel® Xeon® Scalable processors and up to 56 cores. It comes with up to 24 DIMMs, 12 of which can be NVDIMMs or DCPMMs, and up to 7.68TB of memory.[8] With the PowerEdge R640 you can create an NVMe cache pool and use either 2.5" or 3.5" drives for data storage. Create an optimal edge node configuration with scalable processors, memory, and storage.

Award

IT Brand Pulse: 2019 Market Leader, Rackmount Servers: "IT Pros have once again voted Dell EMC the Market leader for Rackmount Servers, making it their third time in seven years."

PowerEdge Servers for Your Predictive Analytics Needs

Worker Nodes

PowerEdge R740

The PowerEdge R740 is an analytics-suited platform, with highly expandable memory (up to 7.68TB) that makes running Spark a breeze. Scale to meet capacity demands with up to 16 x 2.5" or 8 x 3.5" drives and up to 128TB of storage to keep your Hadoop clusters performing optimally. The R740 features the 2nd Generation Intel® Xeon® processor scalable family with up to 56 cores. That's power you will need for processing vast amounts of data.

Award

IDEA 2018 Bronze Award

PowerEdge R740xd

In addition to the PowerEdge R740's capabilities, the Dell EMC PowerEdge R740xd adds extraordinary storage capacity options, making it well-suited for data-intensive applications that require significant storage like Hadoop. Choose up to 24 NVMe drives, or a total of 32 x 2.5" or 18 x 3.5" drives and up to 288TB of storage. The R740xd with 2nd Generation Intel® Xeon® Scalable processors drove 900% more query sets in recent testing than our previous generation server, the R720xd, and delivered the data 99.8% faster. Compared to our most recent version of this server, the R740xd with 2nd Generation Intel® Xeon® Scalable processors hosted 400% more query sets and delivered the data 27% faster.[9]

Review

ServeTheHome: Dell EMC PowerEdge R740xd Review

Worker Node Options, Compared	 Option 1: R740	 Option 2: R740xd
Processing	Up to 56 cores	Up to 56 cores
Storage	128 TB	288 TB
Memory	7.68 TB	7.68 TB

Ready Architectures for Hadoop

Expertise and infrastructure make a difference when building a Hadoop environment.[10] You have the option to reduce effort on your part and purchase a validated blueprint of how to build out your clusters. Ready Architectures for Hadoop are designed to address data analytics requirements, reduce development costs, and enhance performance. While there are many cases of organizations falling behind on deadlines and struggling to make decisions in setting up their Hadoop platforms, that doesn't have to be your company's story.[11]

Dell EMC Ready Architectures for Hadoop are solutions created to meet all your data analytics needs. Dell EMC started building custom Hadoop solutions in 2009 and has the expertise, tools, and solutions needed to drive a successful Hadoop deployment. The Cloudera Hadoop design delivers the key elements of Hadoop within a solution based on Cloudera Enterprise software and Dell EMC hardware, with service options for your convenience. Key benefits include:[12]

- Leveraging an already optimized and validated solution
- Overcome lack of in-house Hadoop expertise
- Shorten project timelines and reduce production issues

To learn more about Dell Ready Architectures for Hadoop, visit our site or contact us today.

Hadoop Migration Services

Ready to take the plunge and migrate to Hadoop, but not sure how to go about it? Dell Technologies Consulting Services can help. You're not alone if you're struggling to manage the opportunities created by data analytics and the heavy workloads associated with vast quantities of data. Dell Technologies Consulting Services offers everything from a data-only migration to a full platform migration, depending upon your organization's needs. It all starts with a current state assessment, leading to a future state solution design. We then pilot the migration to test and validate a sample environment. Following a successful pilot, the solution is fully implemented.[13]

To get started, learn more about Hadoop migration services or contact a Dell Technologies Services Expert.

Customer Story: Epsilon

Let's look at predictive analytics at work. Epsilon is a marketing leader behind loyalty, email, and other marketing programs for world-class brands. It uses a multitude of data points to customize personalized messages for its many recipients, ensuring maximum results for clients. As Epsilon CIO Robert Walden puts it, "Everything we do is centered around data and our ability to get the right message to the right person at the right time."

Speed is paramount for the company to continue its current trajectory of doubling in growth year over year. This requires an infrastructure and a staff that can keep pace. Epsilon automates all its intake processes in the deployment of its Dell EMC PowerEdge R740xd servers. The R740xd provides the flexibility, scalability, and performance to meet the requirements of this demanding business. Running workloads such as Hadoop to apply AI and machine learning to its email customization processes ensures that Epsilon can deliver on its promises.

Epsilon CIO Robert Walden says it best: "We're laser focused on the success of our clients. Everything we do within our infrastructure and our application environments is to facilitate that success. The more effective our applications, solutions, services, infrastructure, and hardware, the better we're able to do that."

Read the Epsilon case study.

Watch the Epsilon case study video.

Epsilon at work

Robert Walden, Epsilon CIO

Conclusion

With modern, automated servers featuring high-powered processors and generous memory and storage options, your organization will be positioned to make the most of the many opportunities generated by predictive analytics. At the same time, mastering the ins and outs of predictive analytics is key to maintaining IT's relevance in your organization. Partner with Dell EMC for solutions that just make sense and take the guesswork out of getting started with machine learning and predictive analytics.

To learn more, contact a Dell EMC sales rep or visit DellEMC.com/Servers.

[1] https://www.dellemc.com/en-us/collaterals/unauth/analyst-reports/products/servers/esg-three-transformational-compute-technologies-verified-to-accelerate-ai-and-business-value-en.pdf

[2] ibid.

[3] https://spark.apache.org/

[4] https://www.dellemc.com/resources/en-us/asset/analyst-reports/products/storage/forrester-delivering-outcomes-by-automating-compute-infrastructure.pdf

[5] https://spark.apache.org/docs/latest/hardware-provisioning.html

[6] ibid.

[7] https://blog.dellemc.com/en-us/ride-ansible-revolution-dell-emc-openmanage/

[8] https://www.dell.com/en-us/work/shop/povw/poweredge-r640

[9] https://www.principledtechnologies.com/Dell/PowerEdge-R740xd-analytics-comparison-0719.pdf

[10] https://www.dellemc.com/en-us/solutions/data-analytics/hadoop/index.htm

[11] https://blog.dellemc.com/en-us/dell-emc-can-help-dive-successful-hadoop-projects/

[12] https://www.dellemc.com/en-us/collaterals/auth/white-papers/products/ready-solutions/Ready_Bundles_for_Hadoop_-_Solution_Overview_China.PDF

[13] https://www.dellemc.com/en-us/collaterals/unauth/offering-overview-documents/services/H16645_Big_Data_Migration_for_Hadoop_svo.pdf