Validated Designs for Analytics
Solution Brief

# Speed Up Large-scale Batch and Streaming Data Processing

## Apache Spark on Kubernetes on optimized Dell EMC Infrastructure

Organizations in a wide variety of industries, ranging from manufacturing and retail to healthcare and finance, can leverage this Spark and Kubernetes-powered analytics design.

The ability to process a large amount of data, be it batch or streaming data, is a must-have for organizations today that want to leverage analytics to drive better business decision-making and power the next generation of machine learning (ML) applications. Traditionally, Apache® Hadoop®, an open-source framework that allows for the distributed processing of large data sets served to fill this need.

Apache Spark® built upon much of the success of Hadoop and took it a step further as a unified analytics engine that performs much of the processing in memory instead of on disk. Claiming a speed improvement of 100X, Spark offers additional benefits over Hadoop such as a built-in ML library and fault-tolerance capabilities.

Spark started with a concept familiar to everyone that works in data science — the data frame. An approach was then devised to distribute it across many systems. This took advantage of the combined memory and computing cores so that data scientists did not have to change the way they traditionally work with data.

Data science meanwhile has expanded beyond the computing power of a single machine to escape the memory and computing core limits of ubiquitous, inexpensive x86 systems. Managing massive, distributed systems that handle the scaling need of data scientists working with increasing data volumes presents new challenges for architects and IT operations professionals. This is where Kubernetes makes a significant contribution.

The relationship between Spark and Kubernetes is conceptually simple. Data scientists want to run many Spark processes that are distributed across multiple systems to have access to more memory and computing cores. Using container virtualization like Docker® to host those Spark processes, Kubernetes orchestrates the creation, placement and lifecycle management of those processes across a cluster of x86 servers.

## Validated Design

The Dell Technologies Validated Design for Analytics — Spark on Kubernetes offers a tested, validated design that describes system building blocks for leveraging the growing capabilities of Kubernetes to manage infrastructure for Spark analytics.

Spark can be run under a number of different Kubernetes distributions. Some of these distributions include Pivotal® Container Service (PKS), Docker Enterprise, Rancher Kubernetes Engine, Ubuntu® Kubernetes, and Red Hat® OpenShift® Container Platform.

This Validated Design uses Red Hat OpenShift Container Platform as its reference platform. Red Hat OpenShift Container Platform is an enterprise Kubernetes platform that incorporates CRI-O as the container engine.

## Resources

- Review the design guide.
- Explore the HPC & AI Innovation Lab.

## Learn more

- Validated Designs for Analytics

## Design components

| Servers | Networking | Software | Storage |
|---|---|---|---|
| Dell EMC PowerEdge R640 | Mellanox® ConnectX-4 Lx dual-port 25GbE SFP 28 rNDC | • Apache Spark<br>• Red Hat OpenShift Container Platform | 2x 800GB SSD SAS mix use 12Gbps and 2.5 in. hot-plug AG drive, 3 DWPD, 4380 TBW |

## Summary

The Spark on Kubernetes design is purpose-built and allows data scientists and data engineers to collaborate to build a full analytics pipeline without having to go outside the Spark ecosystem for data ingestion, data cleansing, data merging, model training and API development for inferencing.

The Validated Design includes a demonstration of Jupyter® notebooks to enable rapid prototyping and visualization capabilities to the data science team. This method uses the same container or Kubernetes management tool set as all the other Spark-specific services.

Finally, this design offers Dell Technologies infrastructure guidance for general-purpose analytics involving all stages of an analytics pipeline using Apache Spark and Kubernetes.