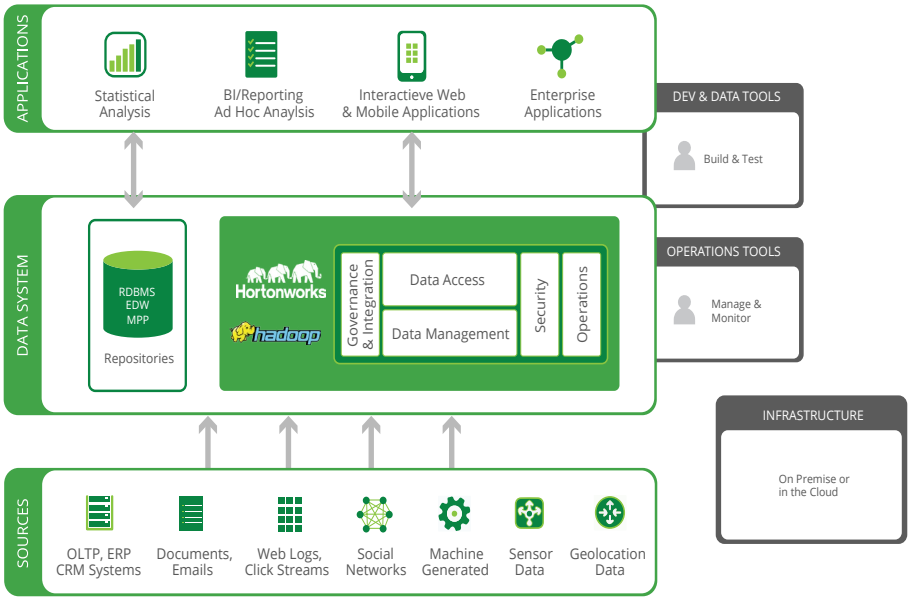
Modern Data Analytics with Big Data Meets Modern Data Center Infrastructures

Hortonworks

Hadoop has clearly become the leading platform for Big Data analytics today. Hortonworks Data Platform (HDP) is a 100% open source, enterprise grade Hadoop distribution driven by Hortonworks. Hadoop represents a modern data architecture designed for deployment on lower cost, high capacity infrastructure that integrates with traditional enterprise infrastructure like RDBMS and MPP systems. Hadoop as a Service (HaaS) is new to the industry but is gaining momentum, leading to a new segment of HaaS providers. This presents an outstanding opportunity for overwhelmed data center admins that need to incorporate Hadoop but don't have the in-house resources or expertise to do so.

HDP is being adopted widely by cloud providers, service providers and enterprises across a wide variety of use cases, including enterprise data warehousing, extract-transform-load (ETL), log processing and compliance.



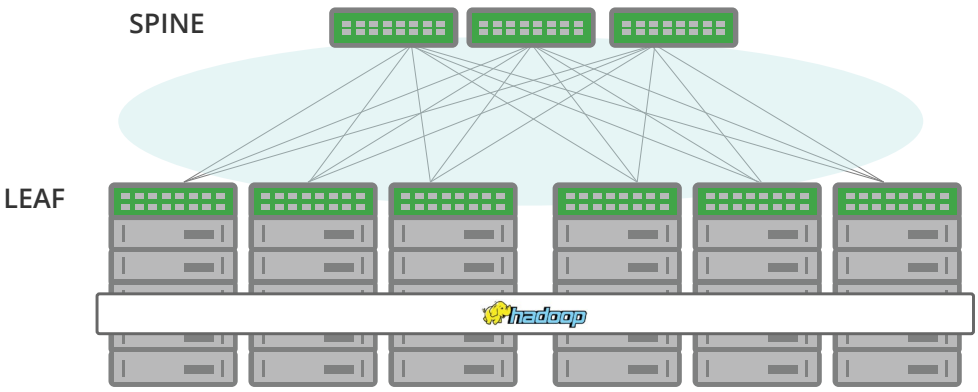
Cumulus Linux

Traditional switches are black boxes with integrated data planes, control planes and feature sets from a single vendor. Although various levels of programmability are available, users are essentially locked in to the functionality provided by these vertically integrated systems and lack the ability to leverage the power of a rich hardware and software ecosystem for specific application requirements and to get sufficient capacity at a reasonable cost. The missing piece was a powerful, reliable, and proven operating system, one that leverages scale and collaboration to enable scores of applications for networking infrastructure.

Cumulus Linux is the first true native Linux operating system built for industry-standard bare metal switches. Being Linux, a Cumulus Linux-powered bare metal switch runs like any other Debian system you've ever used. There is no special command line interface or shell dedicated to switching, routing and other network operations. Cumulus Linux transparently accelerates the Linux data plane while preserving a pure Linux control plane and empowering the wealth of orchestration and management tools that already manage your servers, virtual machines and other appliances.

Cumulus Networks has made affordable the building of high-capacity networks, coupled it with the same management model that the compute nodes have and with a rich ecosystem because they're both based on the same operating system, Linux.

The figure shows a leaf-spine architecture of the underlying network that helps deploy the HDP nodes without any performance penalties in a scalable and agile manner using Cumulus Linux.



- ECMP-based fat tree
- East/west bandwidth
- Better failure handling
- Scalable: Eliminate bottlenecks
- Simple feature set
- Open protocols!

Business Benefits

- **Common philosophy:** Hadoop is a software stack running on disaggregated hardware, usually on Linux-based compute nodes with local disks. Cumulus Linux is very complementary to this approach and adopting it as part of a Big Data deployment yields similar benefits around eliminating vendor lock-in, reducing maintenance costs along with simplifying manageability and increasing availability.
- **CapEx savings:** In adopting Hadoop, customers are often moving away from diskless blades powered by a dedicated storage tier and adopting rack-mounted servers with local storage, often leading to significantly lower CapEx. Such a major departure from existing IT practices also opens up the opportunity to explore cost-effective networking choices driven by the Cumulus Networks hardware ecosystem.
- **OpEx savings:** By combining compute nodes with local disk-based storage, Hadoop helps reduce storage OpEx costs. Similarly, you can expect significant savings on networking OpEx as well, where higher efficiencies can be achieved through automation of all the underlying network infrastructure.
- **Accelerated business growth:** With the ability to customize, simplify and leverage open standards working with Cumulus Linux, customers can plan for architectures to scale without performance bottlenecks or blockers from the underlying network, thus helping build large clusters with the desired flexibility and time to market.

Technical Benefits

- **Automated cluster deployment:** Cumulus Linux, being Linux, provides for the use of a variety of automation toolsets such as Ansible to centrally provision, configure and install with zero touch provisioning capabilities provided by the ONIE (Open Network Install Environment) framework. Big Data users install open source solutions like Zookeeper and Ambari to deploy Hadoop clusters. Several innovations such as networking plug-ins with Ambari APIs can also be easily built in the future.
- **Fast east-west traffic:** Full bandwidth is available between any pair of servers. Hadoop clusters typically have high rates of east-west traffic, which is optimized by the leaf-spine fabric. You can achieve faster recovery with plenty of core bandwidth to re-replicate after failure.

Solution

Data scientists spend significant amounts of time manipulating data, integrating data sets and applying statistical analyses. These types of users typically desire a functionally rich and powerful environment. Ideally, data scientists should have the ability to run Hadoop YARN jobs through Hive, Pig, R, Mahout and other data science tools. Compute operations should be immediately available when the data scientist logs into the service to begin work. Delays caused from starting clusters and reloading data are inefficient and unnecessary. Specifically, a new generation of clustering applications such as Apache Hadoop were based on the fundamental principle that individual nodes and racks can fail rather than assume an infrastructure that isn't error-prone. These applications scaled from a few nodes to tens of thousands of nodes.

Hadoop is a leading example of a modern data storage and processing platform, and is ideally deployed in conjunction with a modern data center infrastructure. In order to best leverage the scale-out capabilities available in Hadoop, compute and storage resources should be deployed using a high performance, non-blocking L3 Clos-based leaf-spine network fabric, which Cumulus Networks provides. Data center infrastructure built entirely using reliable off-the-shelf hardware components is highly cost effective across both CapEx and OpEx.

- **Rack or row-level awareness:** This is typically achieved via a configuration script at the resource manager. Location becomes irrelevant for performance when you leverage a leaf-spine topology and tools like Prescriptive Topology Manager (PTM), which provide a complete blueprint of all physical connectivity to eliminate the issues driven from manual cabling or unreachability concerns.
- **NameNode high availability:** The single master design in HDFS made the NameNode a single point of failure. Hadoop 2.0 addresses this issue by providing a standby NameNode. In case of failure, the secondary NameNode is accessible from the rest of the cluster with little to no loss of bandwidth. You can again leverage the PTM daemon in Cumulus Linux to extract a rack awareness topology through a simple script for the NameNode.
- **Efficiency across wider YARN application support:** In Hadoop 2.0, YARN is able to host non-MapReduce environments as well as MapReduce. Apache Spark is another great analysis framework that is supported. Since Hadoop is moving from being primarily batch focused to incorporating real-time and streaming data processing, a non-blocking network fabric that can provide comprehensive statistics and programmability is beneficial.

Conclusion

Software-driven agility accelerates time to value for Big Data deployments. Intelligent, distributed software architectures ensure that performance can scale in a linear fashion.

- **Performance does not degrade with increasing compute demands with a network fabric based on layer 3 (leaf-spine) Clos architectures. With Cumulus Linux, the networks are more affordable and easier to manage than the legacy network infrastructure. This is because the network, like the application, relies on the same open platform, Linux.**
- **Elasticity is a critical and central consideration for HaaS providers. Environments that support both production jobs and ad hoc analysis by data scientists will experience a wide range of mixed workloads. The ability for these services to adjust to varying workloads in an efficient manner with Cumulus Linux based on non-blocking multi-tier designs is crucial for the infrastructure.**
- **With the introduction of Apache YARN, a whole new range of applications such as HBase, Hive and Spark are easily deployed by more than just Web-scale companies. And with projects like Ambari, they make running large clusters manageable. A Cumulus Linux innovation such as PTM enables a Hadoop node to identify rack locality in an easier and more precise way. These are just some of the examples through which Cumulus Linux can strategically enhance Big Data deployments.**

The Apache Hadoop project made this new generation of clustering applications accessible to all. And Cumulus Networks is making the network that Hadoop relies on accessible for all.

Get Started

For more information, visit **www.hortonworks.com** and **cumulusnetworks.com/solutions/Big Data**.

About Cumulus Networks®

Cumulus Networks is bringing the Linux revolution to networking. Founded by veteran networking engineers from Cisco and VMware, Cumulus Networks makes the first Linux operating system for networking hardware and fills a critical gap in realizing the true promise of a software-defined data center. For more information visit **cumulusnetworks.com** or follow us on Twitter **@cumulusnetworks**.

About Horton Works

Founded in 2011 by 24 engineers from the original Yahoo! Hadoop development and operations team, Hortonworks has amassed more Hadoop experience under one roof than any other organization. Hortonworks team members are active participants and leaders in Hadoop development, designing, building and testing the core of the Hadoop platform with years of experience in Hadoop operations, and are best suited to support your mission-critical Hadoop project.

Hortonworks performs all of its development within the processes of the Apache Software Foundation – our code is 100% open source with zero proprietary extensions. Learn more at **www.hortonworks.com** or follow on Twitter **@hortonworks**.

20141009-1C