SOLUTION BRIEF

Delivering Powerful Analytics Using Cloudera Distribution for Hadoop deployed on High-speed Ethernet

EXECUTIVE SUMMARY

The rise of mobile technology and the Internet of Things (IoT) has resulted in sensors in cars, industrial equipment, and medical devices to name a few. Due to these and other similar devices, an abundance of data is being created form areas that have never produced digital data. This data is increasingly unstructured in nature and is typically stored in Hadoop and NoSQL systems. New actionable insights can be gleaned from this data to fuel new business growth and the spoils of this new-found data can spur a competitive advantage for companies that can successful analyze it. However, for big data to deliver on the promise of its vast potential, technology must be in place to enable organizations to not just capture and store data but efficiently transport that data. Providing the most efficient data delivery and acceleration offload technologies allows for organization to capitalize on this data by gaining new insights that can be used to improve business. This solution brief discusses the data analytic advantages of pairing a high-speed Ethernet network with Cloudera to solve problems involving massive amounts of data and computation.

THE CLOUDERA DATA PLATFORM

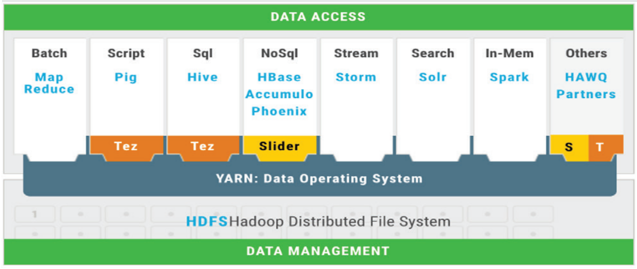
Cloudera delivers the modern platform for machine learning and advanced analytics built on the sates open source technologies. As a leading innovator of Hadoop and Apache Spark, Cloudera delivers an enterprise-ready open data platform. Cloudera Enterprise Data Hub is built for modern big data analytic applications and uses the Hadoop Distributed File System (HDFS) for scalable, fault-tolerant big data storage and Hadoop's centralized Yet Another Resource Negotiator (YARN) architecture for resource and workload management. YARN enables a range of data processing engines including SQL, real-time streaming and batch processing, among others, to interact simultaneously with shared datasets, avoiding unnecessary and costly data silos and unlocking an entirely new approach to analytics. This allows for an open platform that works perfect with low-cost commodity compute and storage servers for running big data workloads while providing which allows for de-coupling of compute and storage to enable optimized configurations.

SOLUTION HIGHLIGHTS

- Conduct analysis and formulate new hypotheses quicker than before
- Continuous operation with Zero-Touch deployment
- Accelerate Hadoop deployment with enterprise grade reliability from Mellanox
- Lowest TCO by maximizing system resources and supporting multiple workloads with the most economical choice for building a Hadoop cluster

Increasing Hadoop Efficiency

10/25/40/50/100G Ethernet speeds, sub-microsecond latency and offloading mechanisms such as RDMA, Erasure Coding, TCP, UDP, as well as overlay network and OVS offloads to free CPU resources.

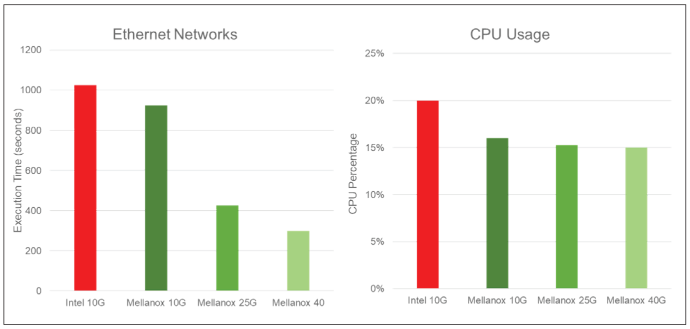


HIGHLY SCALABLE PLATFORM OPTIMIZED FOR PERFORMANCE

The high-performance data platform from Cloudera enables business analysts to formulate new business values quicker than before using different big data applications. The workloads could be SQL-based (using Hive, Impala, Spark SQL, etc.), the newer class of NoSQL-based, HDFS or the more advanced applications using machine learning, graph, and streaming analytics. Due to the advantages of HDFS and the increasing usage of flash storage in production grade Hadoop environments, network bandwidth often becomes the bottleneck during data ingestion and analytics. Investing in faster servers and flash storage for Cloudera does not make sense if performance is restricted by the network. Mellanox provides the most efficient end-to-end Ethernet network tailored for big data applications at 10/25/40/50 and 100G Ethernet speeds.

CLOUDERA AND MELLANOX, BETTER TOGETHER

With industry-leading performance and IT efficiency combined with the best of open innovation to accelerate big data analytics. Mellanox Spectrum® Ethernet switches feature consistently low latency and can support a variety of non-blocking, lossless fabric designs. The Mellanox's ConnectX® adapters reduce the CPU overhead in packet processing through advanced hardware-based stateless offloads and flow steering engines. This allows big data applications utilizing TCP or UDP over IP transport to achieve the highest throughput and application density. These advanced offloads reduce CPU overhead in IP packet processing, allowing completion of heavier analytic workloads in less time for big data clusters so organizations can unlock and scale data-driven insights for their business like never before.



EASE OF END-TO-END MANAGEMENT

In large scale-out software ecosystems such as Hadoop and NoSQL, delivering high availability and low downtime becomes a crucial challenge for administrators when deploying a wider ecosystem of big data software such as Spark, Drill, and Sqoop, in addition to MapReduce applications. Cloudera adds to the ease of management with capabilities with capabilities that simplify business continuity in Hadoop distributions, tiered storage, security, and provisioning. With Cloudera Manager, the Hadoop administration tool, administrators can manage and monitor the Cloudera deployment to ensure a reliable, high performance system. Mellanox Onyx™ provides the most advanced network orchestration, automation and monitoring platform for IT administrators to get a 360-degree view of the entire network. Furthermore, with the One-Click feature in Onyx, data center administrators can easily configure their large scale-out network in simple templates thereby allowing them to automate replication of the network state whenever new data nodes are added to the cluster.

With its industry-leading core count, EPYC can enable more VMs and more robustly configured VMs per server than previously possible. EPYC natively supports up to 32 NVMe or SATA devices in both 1 socket and 2 socket designs, enabling streamlined Software Defined and Direct Attached Storage solutions.

CONCLUSION

With the right technologies and practices, big data analytics can help organization make rapid strides in gathering comprehensive knowledge about customers and use predictive insights to improve satisfaction, loyalty and ultimately, profitability. Although Apache Hadoop offers a powerful tool for analyzing large and diverse data sets, deploying a successful, reliable and high-performance infrastructure can be a daunting challenge. The Cloudera distribution for Hadoop provides critical technology advances to make Hadoop implementation easy, dependable, and fast for production deployments. By taking advantages of Mellanox's lossless switches and advanced offloads and stateless engines with Mellanox adapters, IT organizations no longer have to choose between performance and reliability. With integrated capabilities and a high-performance network, businesses can achieve the competitive advantages of big data analytics faster with less risk and with the confidence of an enterprise grade ecosystem.

About Mellanox

Mellanox Technologies is a leading supplier of end-to-end InfiniBand and Ethernet interconnect solutions and services for servers and storage. Mellanox interconnect solutions increase data center efficiency by providing the highest throughput and lowest latency, delivering data faster to applications and unlocking system performance capability. Mellanox offers a choice of fast interconnect products: adapters, switches, software, cables and silicon that accelerate application runtime and maximize business results for a wide range of markets including high-performance computing, enterprise data centers, Web 2.0, cloud, storage and financial services.

To find out more, visit our website: www.mellanox.com

350 Oakmead Parkway, Suite 100
Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com