Optimizing Hadoop with Hortonworks and Cisco

Hortonworks Data Platform on the Cisco UCS platform delivers a stable, scalable architecture for open-source Apache Hadoop.

Big data is exploding. And Apache Hadoop is expanding with it as the framework of choice for enterprises seeking to leverage their masses of data for competitive advantage. The most capable, stable, scalable Hadoop deployment, in one certified and supported package, is Hortonworks Data Platform (HDP) on the Cisco® Unified Computing System™ (UCS®) platform.

Challenge: Capitalizing on the Data Explosion

In today's digital market, data is the new competitive edge. That is why smart organizations are storing and analyzing massive quantities of data, looking for the information that will help them improve operations or deliver better products and services. But it is not easy to process large amounts of data—and it gets harder every day, as data volumes grow and legacy systems are overwhelmed.

And the challenge will only get tougher. The emergence of the "Internet of Everything" is greatly intensifying the volume and variety of data. A study by IDC shows the digital universe is doubling in size every two years and predicts it will rise tenfold by 2020. Enterprises that succeed in this era will collect as much data as they can and draw value from it, analyzing it to discover new ways of improving operations or learning about customers. But to do so, they need the best big data solutions available.

> “Working together, Hortonworks Data Platform and Cisco UCS Integrated Infrastructure for Big Data provide an industry-leading platform for Hadoop-based applications.”

Combined Solution: Hortonworks Data Platform and Cisco UCS Integrated Infrastructure for Big Data

Apache Hadoop has enabled enterprises of all sizes to utilize big data cost-effectively. Hadoop is an open-source software framework that allows for the distributed processing of large data sets across clusters of hardware using simple programming models. It is designed for high availability and fault tolerance and can scale from a single server up to thousands of machines.

But Hadoop alone is not enough. Hortonworks Data Platform delivers enterprise Hadoop designed to meet the ever-expanding demands of enterprise data processing. HDP is a platform for multiworkload data processing across an array of processing methods, from batch to interactive to real time, packaged with additional open-source capabilities for governance, integration, security, and operations.

HDP enables customers to unlock the value of big data using Hadoop. The platform is the only 100 percent open-source distribution of Hadoop, developed, tested, and hardened with enterprise rigor and delivered simultaneously on Linux and Microsoft Windows. It is supported by enterprise-grade deployment, training, and technical services, as well as an extensive ecosystem of platform and solution partners.

Challenges

- **The estimated 2.8 ZB of data produced in 2012 is expected to grow to 40 ZB by 2020. According to IDC, 85 percent of this increase will consist of new data types, with machine-generated data projected to increase by 15 times by 2020.**
- **More and more of today's data has little or no structure—or structure that changes too frequently for reliable schema creation at the time of collection.**
- **Incoming data may have little or no value as individual records or small groups of records. But high volumes and longer historical perspectives can be mined for patterns and leveraged in advanced analytic applications.**

Key Features and Benefits of HDP

✓ Cohesive integration

Traditional solutions are not always well-suited to processing nontraditional data sets, such as text, images, machine data, and online data. The HDP-Hadoop solution enables enterprises to incorporate both structured and unstructured data in one data-management system.

✓ Easy online archiving

It is not always obvious what stored data will be valuable in the future, so enterprises may not be able to justify expensive processes to capture, cleanse, and store large amounts of data. This is not a problem with HDP. It scales easily, so data can be stored for years without significant incremental costs.

✓ Ready access

Data is not useful if it is not accessible. Hadoop clusters are a low-cost solution for storing massive data sets that are easily available. Hadoop effectively scans all data and is complementary to databases that are efficient at finding subsets of data.

✓ Flexible deployment

HDP offers the broadest range of deployment options for Hadoop, from Windows Server to Linux to virtualized cloud deployments. It is also the most portable Hadoop distribution available, allowing easy and reliable migration from one deployment type to another.

With Hadoop on HDP, enterprises can easily and economically store, manage, and process large data sets, as HDP combines the most effective and stable versions of Hadoop into a single tested and certified package. The HDP-Hadoop solution is always up to date because Hortonworks continually delivers the latest innovations from the open-source community, along with the testing and quality that organizations expect from enterprise-quality software.

The tremendous competitive implications of big data demand that it be treated as a long-term initiative. That means enterprises must think long-term about what infrastructure platform their big data is running on. With billions of people and devices producing data—and more coming every day—enterprises need a big data platform that can scale to handle massive quantities of data, speed deployment of new resources, efficiently manage those resources to lower cost of ownership, and deliver outstanding performance for years to come. Only Cisco UCS Integrated Infrastructure for Big Data delivers all these essentials.

Business Results: Scalable, Effective Data Processing at Low Cost

Hortonworks Data Platform on the Cisco UCS platform delivers massive scalability. So even as the volume and variety of their data rapidly expand, enterprises can be confident they'll be able to store and manage their data flow. The Cisco UCS fabric-based architecture uniquely integrates server, network, and storage. This highly efficient infrastructure lets businesses manage up to 10,000 UCS servers as if they were a single pool of resources so they can support the largest data clusters. And, as an enterprise's big data deployments grow in size, the solution's use of Cisco Unified Fabric technology significantly lowers costs by reducing the number of switches and cables they will require.

Enterprises will need to be able to capture intelligence from both data at rest in the data center and real-time data at the edge of the network. The broad portfolio of the Cisco UCS solution provides the flexibility to process data where it makes the most sense. The Cisco UCS C240 M4 rack server is extremely popular for Hadoop-based big data deployments at the data center core. Cisco UCS Mini is an all-in-one solution that's ideal for processing data at the edge, delivering servers, storage, and networking in an easy-to-deploy, compact form factor.

The UCS platform's integrated design also delivers outstanding performance and enables powerful management-automation capabilities. For instance, Cisco UCS Manager abstracts all configuration, identity, and I/O connectivity information into a UCS service profile that can be applied to other servers. This intelligent programmability allows you to rapidly and consistently deploy new big data servers, restore a failed server, and update releases across an entire network of UCS servers. Individual UCS servers can be deployed 84 percent faster and with fewer steps than in traditional environments, with automation freeing staff from tedious, time-consuming chores that can also be the source of errors. This capability makes the entire data center more cost-effective, resulting in a 51 percent reduction in management costs.

Solution

- **Fully tested and certified, the Cisco UCS solution for HDP is based on UCS Integrated Infrastructure for Big Data, a highly scalable and efficient architecture designed to meet the demands of rapidly growing big data environments with cohesive data integration and powerful management-automation capabilities.**
- **With Cisco UCS Integrated Infrastructure for Big Data, enterprises can smoothly integrate Hadoop with their computing, networking, and storage resources to run Hadoop clusters.**
- **Together, Cisco and Hortonworks offer full lifecycle support and a portfolio of Cisco Advanced Services to ensure customer success at every step, from development to proof of concept to staging.**

In addition, UCS Director Express for Big Data is integrated with major Hadoop vendors to provide centralized visibility across the entire Hadoop infrastructure. Enterprises can now provision on-demand Hadoop clusters and manage both physical and software infrastructures from a single management pane.

HDP on the Cisco UCS Integrated Infrastructure for Big Data solution also takes full advantage of the benefits of Cisco Application Centric Infrastructure, an innovative architecture that optimizes and accelerates the application-deployment lifecycle. This solution helps IT departments simplify the way they provision the data center resources that are critical to the performance of big data applications such as Hortonworks.

As the number of an enterprise's big data workloads increases, the enterprise network will play a more important role in ensuring that workloads are completed and insights delivered on a timely basis. Cisco Application Centric Infrastructure applies network intelligence to dynamically load-balance big data flows across racks in a multirack big data cluster. For instance, Application Centric Infrastructure can detect congestion on the top-of-rack switches and reroute big data traffic using alternate pathways—potentially resulting in triple-digit-percentage improvements in the workload completion times and throughput of big data jobs. Cisco Application Centric Infrastructure also provides the network with workload awareness. So the programmable infrastructure can prioritize the small packets associated with big data traffic traveling between compute nodes ahead of larger packets associated with other workloads that could otherwise delay big data jobs. Consequently, an enterprise can be confident that its mission-critical big data jobs complete much faster.

On the security front, HDP delivers protection at every layer of the Hadoop stack, while Cisco offers the industry's most extensive portfolio of integrated advanced malware protection solutions. With Cisco Advanced Malware Protection, customers get continuous visibility and control to defeat malware across the extended network and the full attack continuum—before, during, and after an attack.

Next Steps

To find out more about how Cisco and Hortonworks big data solutions can benefit your organization, visit http://hortonworks.com/partner/cisco/.

Business Results

- **Enterprises can quickly and efficiently scale to 10,000 UCS servers to support the largest data clusters.**
- **With Cisco UCS Manager, new resources can be deployed 84 percent faster and with fewer steps than in traditional environments.**
- **Powerful Cisco management automation results in a 51 percent reduction in management costs compared to traditional systems.**