DELLEMC

HADOOP SOLUTION USING DELL EMC ISILON AND CLOUDERA ENTERPRISE

Efficient, Flexible In-Place Hadoop Analytics

THE BIG DATA OPPORTUNITY

The rapid growth of data represents a significant challenge for many enterprises across a wide range of industries today. Many organizations are realizing that Big Data is a valuable asset that can be leveraged to uncover new opportunities to accelerate their businesses and gain a competitive advantage. Central to realizing this opportunity is Hadoop, an innovative Big Data analytics engine designed specifically to analyze large- scale datasets.
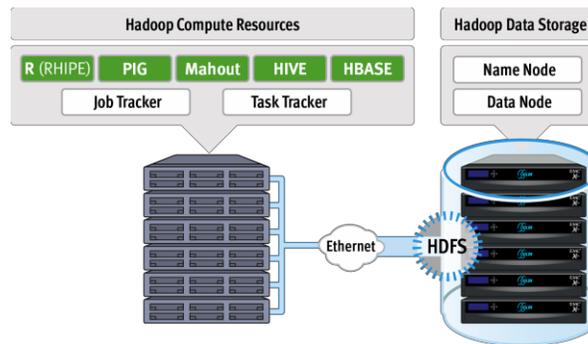
CHALLENGES WITH TRADITIONAL HADOOP DEPLOYMENTS

While most companies start out using Hadoop to capture and store data for future analysis in one consolidated pool, many already own network-attached storage (NAS) - or storage area network (SAN)-based storage systems existing in silos across the organization. How do you utilize these data storage systems without the required separate capital investment and added management resources? How do organizations work around:

- Inefficient storage with poor utilization and difficult management due to silos

- Manual ingest of large datasets into Hadoop—time and resource consuming

- Accessing or sharing data and analytics results across the organization as most systems do not support Hadoop Distributed File System (HDFS)

COMBINED ISILON AND CLOUDERA APPROACH

Dell EMC® Isilon® together with Cloudera provide a complete, tested, and popular solution encompassing Apache Hadoop distribution and related projects. The combination of Isilon Big Data shared storage and the Cloudera Enterprise data management platform helps organizations speed up time to insights, enforce consistent security, enable multiprotocol access, and eliminate storage silos with a powerful yet simple, efficient, scalable, and complete solution.

Isilon is the first and only scale-out NAS platform with native HDFS support in addition to traditional Server Message Block (SMB), network file system (NFS), HTTP, and FTP. This enables
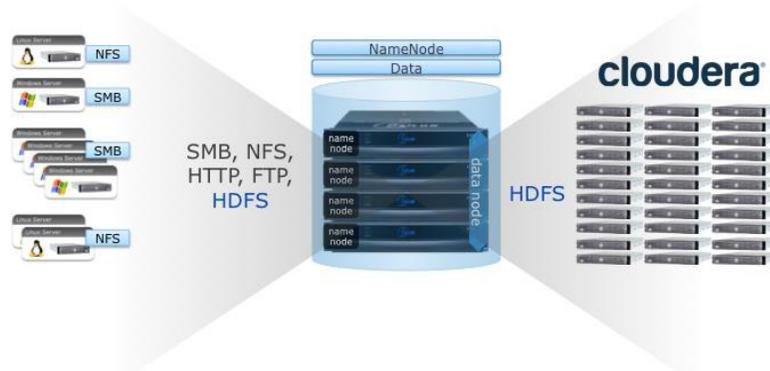
organizations to deploy one shared storage system that works across traditional and new and emerging workloads.

In addition to delivering the core elements of Hadoop, Cloudera Enterprise focuses on making the Big Data management platform secure, managed, governed, and open. Cloudera Enterprise is a thoroughly tested, documented, and supported solution that takes the guesswork out of building the Hadoop deployment.

Together Isilon and Cloudera provide a comprehensive scalable storage and distributed computing solution to meet most enterprise Big Data analytics needs.

Dell EMC ISILON SHARED STORAGE

Isilon combines a powerful yet simple, highly efficient, and massively scalable storage platform with integrated support for Hadoop analytics. Isilon native support for the HDFS allows you to quickly implement an in-place data analytics solution and avoid unnecessary capital expenditures, increased operational costs, and time-consuming replication of your Big Data to a separate infrastructure. Simply connect your analytics compute resources to your Isilon storage system, and you're ready to begin your analytics projects immediately.

GAIN FASTER TIME TO INSIGHTS

Isilon's in-place data analytics approach allows you to eliminate the time and resources required to replicate your Big Data set into a separate Hadoop infrastructure. For example, it can take over 24 hours to copy 100 TB of data over a 10 Gb line. Instead, with Isilon, you can initiate data analytics projects immediately and get to results quickly.

ENJOY INCREASED FLEXIBILITY

Isilon supports multiple instances and multiple versions of Apache Hadoop distributions simultaneously. This allows you to leverage the tools you need for each of your unstructured data analytics projects. In addition to native HDFS 1.0 and HDFS 2.0 support, Isilon solutions include integrated support for a wide range of industry- standard protocols, including NFS, SMB, HTTP, FTP, SWIFT, and REST-based object access.

GAIN MASSIVE SCALABILITY

With Isilon, you can have massive room for growth for your unstructured data assets and related analytics projects. Isilon scales from 18 TB to 68 PB of capacity in a single Isilon cluster. The Dell EMC Isilon OneFS® operating system allows a storage system to grow symmetrically or independently as more space or processing power is required. This provides a true grow-as-

you-go approach and the ability to scale out as business needs dictate. With Isilon, you can scale capacity and performance.

PROTECT YOUR BIG DATA ASSETS

Isilon provides unsurpassed levels of data protection and availability to meet a wide range of enterprise needs. OneFS enables all nodes in the Isilon storage cluster to become, NameNodes, thus improving the resiliency of your Hadoop environment.

Isilon also offers end-to-end data protection options for fast and efficient data backup and recovery. You can schedule snapshots as frequently as needed to meet your specific recovery point objectives. For reliable disaster recovery protection, Isilon provides fast data replication, along with push-button failover and failback simplicity, to further increase the availability of your data assets. We also automatically route clients to nodes based on load, this prevents over subscribing already heavily loaded storage nodes with high workload. Furthermore we offer traffic segregation with the use of Access Zone so that Hadoop workloads can be isolated from other critical workloads.

SECURE YOUR BIG DATA ASSETS

To help you meet regulatory compliance and corporate governance requirements, Isilon offers robust security options, including file system auditing and write once, read many (WORM) data protection to prevent accidental or malicious alteration or deletion. With Isilon, you can also provide secure role separation between storage administration and file system access, as well as authentication zones, to create secure, isolated storage pools for specific departments within your organization. Whether your data is in-flight or at rest, we offer an end to end encryption solution.
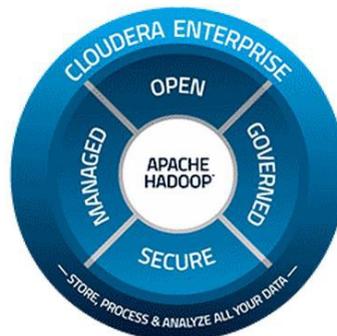
IMPROVE STORAGE UTILIZATION

With Isilon, you can consolidate your storage infrastructure, including semistructured and unstructured data assets. Isilon provides a storage utilization rate of over 80 percent, which can be further improved by up to 35 percent with the use of Dell EMC Isilon SmartDedupe data deduplication that eliminates redundant data. You need less storage capacity and physical space to house your dataset—reducing both initial capital outlay and ongoing operating costs.

CLOUDERA ENTERPRISE

- Unified—one integrated system, bringing diverse users and application workloads to one pool of data on a common infrastructure, with no data movement required

- Secure—perimeter security, authentication, granular authorization, and data protection

- Managed—native high-availability, fault-tolerance, and self-healing storage; automated backup and disaster recovery; and advanced system and data management

- Open—Apache-licensed open source to ensure that your data and applications remain yours, and an open platform connects with all of your existing investments in technology and skills

CLOUDERA ENTERPRISE

Cloudera Enterprise helps you become information driven by leveraging the best of the open source community with the enterprise capabilities you need to succeed with Apache Hadoop in your organization. Designed specifically for mission-critical environments, Cloudera Enterprise includes CDH, the world's most popular open source Hadoop-based platform, as well as advanced system management and data management tools, plus dedicated support and community advocacy from our world- class team of Hadoop developers and experts. Cloudera is your partner on the path to Big Data.

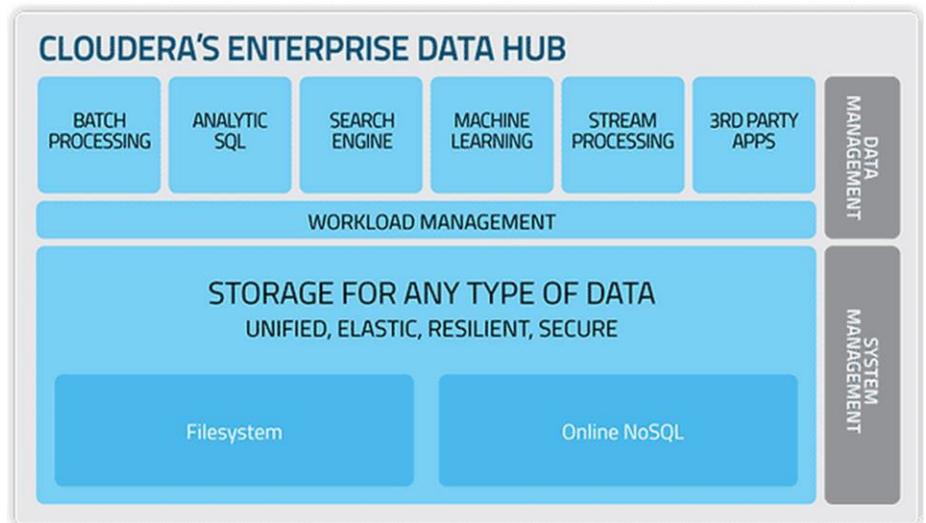Dell EMC Isilon & Cloudera CDH Solution Overview September 2016

RETHINK DATA MANAGEMENT

Cloudera Enterprise is designed to function as an enterprise data hub. It is:

- One massively scalable platform to store any amount or type of data, in its original form, for as long as desired or required

- Integrated with your existing infrastructure and tools

- Flexible to run a variety of enterprise workloads—including batch processing, interactive SQL, enterprise search, and advanced analytics

- Robust security, governance, data protection, and management that enterprises require

With Cloudera Enterprise, today's leading organizations can put their data at the center of their operations to increase business visibility and reduce costs while successfully managing risk and compliance requirements.

ONLINE NOSQL—HBASE

HBase is a distributed key-value store that helps you build real-time applications on massive tables with billions of rows and millions of columns with fast random access. When you deploy HBase as part of Cloudera Enterprise Flex Edition or Data Hub Edition as part of an enterprise data hub, you can rely on our market-leading technical support for HBase, as well as actively influence the future of the project.

ANALYTIC SQL—CLOUDERA IMPALA

Cloudera Impala is the industry's leading massively parallel processing (MPP) SQL query engine that runs natively in Apache Hadoop. With Impala, you can enable analysts and data scientists to directly interact with any data stored in Hadoop using their existing business intelligence (BI) tools and skills through an industry-standard SQL interface. You can also offload self-service business intelligence to Hadoop, relieving the burden on existing analytical databases and reducing your "BI backlog."

SEARCH—CLOUDERA SEARCH

Cloudera Search lets your users query and browse data in Hadoop just as they would search Google or their favorite eCommerce site. Powered by Apache Hadoop and Apache Solr, the enterprise standard for open source search, Cloudera Search brings scale and reliability for a new generation of integrated, multiworkload search. Through its unique integrations with Cloudera Enterprise, Cloudera Search gains the same fault tolerance, scale, visibility, security, and flexibility provided to other enterprise data hub workloads.

IN-MEMORY MACHINE LEARNING AND STREAM PROCESSING— APACHE SPARK

Apache Spark (incubating) is an open source, parallel data processing framework that complements Apache Hadoop to make it easy to develop fast, unified Big Data applications combining batch, streaming, and interactive analytics on all your data.

SUMMARY

Isilon together with Cloudera provides a Big Data storage and analytics solution that is powerful yet simple and highly efficient with a massively scalable shared storage platform and industry-leading Hadoop analytics distribution. With Isilon, you can streamline your storage infrastructure by consolidating large-scale file and unstructured data assets, eliminating silos of storage and reducing costs. Isilon solutions also allow you to reduce time to insights with in-place analytics. At the same time, you gain the flexibility to support multiple instances of Apache Hadoop distributions from different vendors simultaneously.

With Cloudera Enterprise, you get the core elements of Hadoop along with additional components such as a user interface, plus necessary enterprise capabilities such as security, and integration with a broad range of hardware and software solutions. All the integration work is done for you, and the entire solution is thoroughly tested and fully documented. By taking the guesswork out of building out your Hadoop deployment, Cloudera Enterprise gives you a streamlined path to success in solving real business problems.

TAKE THE NEXT STEP

Contact your Dell EMC sales representative or authorized reseller to learn more about how Isilon Big Data storage and analytics solutions can benefit your organization.

Also see our solutions in the Dell EMC Store at https://store.emc.com/isilon and Cloudera solutions at www.cloudera.com.