SCALE-OUT NAS FOR GREENPLUM HD

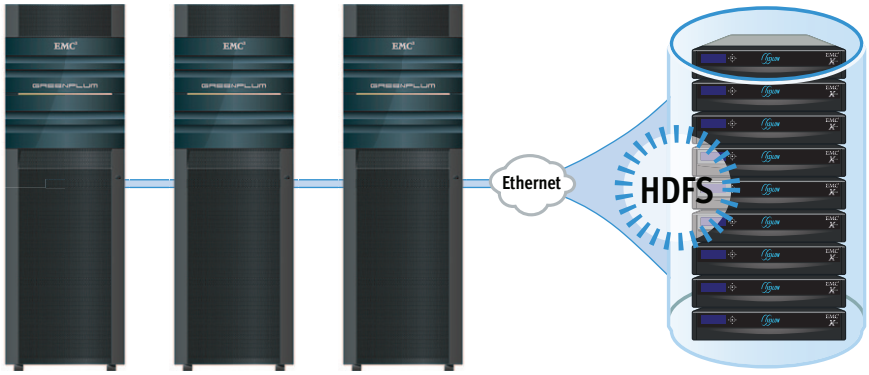
EMC Isilon Big Data Storage and Analytics Solution

ESSENTIALS

- Industry's first and only Enterprise Scale-Out NAS solution for Hadoop workflows
- Combines Greenplum HD and Isilon scale-out NAS
- Greenplum HD DCA Appliance accelerates implementation
- Scale Greenplum HD Compute and Storage resources independently
- Comprehensive Hadoop solution engineered for enterprise requirements:
  - Automatic load balancing
  - Reliable and efficient data protection with snapshot functionality
  - 80% capacity utilization
  - Scale to 85 GB/s and 15 PB in a single file system
  - Flexible management and monitoring to optimize system performance

INTEGRATED HADOOP APPROACH

EMC Isilon scale-out NAS is the first and only Enterprise NAS solution that can natively integrate with the Hadoop Distributed File System (HDFS) layer. By treating HDFS as an over the wire protocol, you can quickly deploy a comprehensive big data analytics solution that combines Greenplum HD with Isilon scale-out NAS storage systems to provide a powerful, highly efficient and flexible data storage and analytics ecosystem. The combined Greenplum HD and Isilon solution enables your organization to avoid the resource-intensive complexity of traditional Hadoop deployments by providing a packaged, yet comprehensive Hadoop system. This approach enables you to focus more on analyzing your business rather than spending valuable resources struggling with the technical complexities of configuring and managing a Hadoop cluster.



Leverage the robust analytic capabilities of Greenplum HD Hadoop workflows with the efficiency, reliability and flexibility of Isilon Scale-Out NAS.

REALIZING THE PROMISE OF HADOOP

The volume of data produced and managed by enterprises continues to expand rapidly, especially in today's internet age. Organizations across all industries are looking to leverage their ever-growing big data assets to identify key trends and new opportunities that can help accelerate their business. As the volume and sources of data have expanded dramatically, traditional techniques to store and analyze this information have not kept pace. Existing analytics architectures have proven to be too expensive and too slow to handle the massive data volumes modern enterprises are facing.

Apache Hadoop has emerged as an innovative big data analytics engine that is designed to support data-intensive distributed applications and reduce the time required for enterprise data analysis especially of unstructured data. Hadoop is an open-source software framework for running large batch-oriented information analytics jobs across clusters of commodity servers.

BENEFITS

- Highly reliable and flexible infrastructure with no single point of failure
- Ability to use HDFS as a true file system
- Efficient and easy-to-manage Hadoop environment
- Multi-protocol access to optimize entire Analytics workflow
- Monitoring of analytics resources with Greenplum Command Center
- Interoperability for structured analytics in Greenplum Database
- Native connectivity for Hadoop and Eco-system components, such as:
  - Hive
  - HBase
  - Pig
  - Mahout

SUPPORTED DISTRIBUTIONS

Greenplum HD 1.1

Hadoop is comprised of two major components; MapReduce and the Hadoop Distributed File System (HDFS). MapReduce is a parallel task processing mechanism that takes a query and runs it in parallel on multiple nodes to more efficiently process massive datasets. HDFS is the distributed file system that a typical Hadoop cluster uses to store all of the data to be analyzed by MapReduce. Hadoop is often utilized with Hadoop components such as HBase, Pig and Hive for expanded database computing capabilities.

Data stored in Hadoop is typically replicated multiple times, and distributed across the Hadoop cluster to optimize performance and reliability. Traditional Hadoop deployments run on a cluster of commodity servers for computation (MapReduce), utilizing direct attached storage for data storage (HDFS) and connected together via a network. A single server referred to as the NameNode, stores all of the metadata for the files stored in HDFS. This traditional approach for Hadoop environments results a number of challenges for enterprises:

- Technical complexities of Hadoop implementation and management
- NameNode is a single point of failure
- Storage utilization of direct attached storage is inefficient, a problem compounded by HDFS replication
- Managing large pools of locally attached disks is complex and expensive
- Inefficient data staging and loading process
- Inability to independently increase performance or capacity
- Complex integrations across multiple open source projects

To address these challenges, our solution combines the power of Greenplum HD with Isilon scale-out NAS to provide you with a comprehensive big data storage and analytics solution that delivers unsurpassed value.

GREENPLUM HD

Greenplum HD enables you to take advantage of big data analytics without the overhead and complexity of a project built from scratch. Available as software or in a pre-configured Data Computing Appliance Module, Greenplum HD provides a complete platform, including installation, training, global support, and value-add beyond simple packaging of the Apache Hadoop distribution. The Greenplum HD Module combines Hadoop and the Greenplum Database in one purpose-built Data Computing Appliance.

Greenplum HD is a 100 percent open-source certified and supported version of the Apache Hadoop stack that includes HDFS, MapReduce, Hive, Pig, Hbase and Zookeeper. Backed by the world's largest Hadoop support organization and tested at scale in Greenplum's 1,000 node Analytics Workbench, Greenplum HD brings flexible storage options to an enterprise-ready Hadoop stack. Greenplum HD makes Hadoop faster, more dependable, and easier to use.

BIG-DATA ANALYTICS WITH EMC ISILON SCALE-OUT NAS

EMC Isilon scale-out NAS provides native integration with HDFS along with the simplicity, efficiency, flexibility and reliability that you need to maximize the value of your Hadoop data storage and analytics workflow investment.

SIMPLICITY

- Simplified storage management with InsightIQ
- Eliminate complexities of managing large pools of direct attached storage
- Scale performance and capacity in less than 60 seconds while servicing Hadoop workflows
- Gain the ability to use HDFS as a true file system
- Simplify entire workflow by loading data over standard protocols

EFFICIENCY

- Bring Industry leading storage efficiency to Hadoop
- EMC Isilon scale-out NAS can provide >80% utilization from a single pool of shared storage. Traditional Hadoop storage deployments using direct attached storage typically require two to three times more capacity due to the inherent inefficiency of DAS.
- Ability to eliminate resource intensive import/export of data in/out of Hadoop
- Achieve reduced data center footprint due to high density storage and Reed Solomon parity striping across a single volume/file system.
- Leverage better Hadoop compute CPU utilization, smaller datacenter footprint and ease of management to keep TCO in check

FLEXIBILITY

- Flexibility to expand Hadoop's capacity or compute independently
- Leverage multiple standard protocols to distribute insight to other components of the big data analytics workflow
- Leverage EMC Isilon scale-out architecture to expand capacity economically
- Scale aggregate performance to 85GB/s out of one single volume/filesystem

RELIABILITY

- Benefit from using load balancing and failover with SmartConnect
- Every EMC Isilon node is a namenode for Hadoop
- Apply robust and efficient distributed data protection at a file/folder level with FlexProtect
- Achieve enterprise class data protection with SnapShotIQ, SyncIQ, SmartLock and SmartConnect

CONTACT US

To learn more about how EMC products, services, and solutions help solve your business and IT challenges contact your local representative or authorized reseller—or visit us at www.EMC.com

EMC², EMC, and the EMC logo, are registered trademarks or trademarks of EMC Corporation in the United States and other countries. All other trademarks used herein are the property of their respective owners. © Copyright 2012 EMC Corporation. All rights reserved. Published in the USA. 01/12 Solution Overview H8319

EMC Corporation
Hopkinton, Massachusetts 01748-9103
1-508-435-1000 In North America 1-866-464-7381
www.EMC.com

EMC²