THE PROMISE OF THE SCALE-OUT DATA LAKE

The rapid growth of unstructured data represents a significant challenge for many enterprises across a wide range of industries today. As the volume and sources of data have expanded dramatically, traditional techniques to store and analyze information have proven to be too expensive and too slow to handle the massive data volumes modern enterprises produce and manage.

Central to realizing this opportunity is an open and flexible modern data architecture, a scale-out data lake, that provides:

- A highly scalable and efficient infrastructure that lowers costs and easily keeps pace with growing data storage requirements
- Powerful yet easy-to-use computing platform and analytic tools to unlock the business value of information that lives in the data
- Enterprise class data protection to maximize availability and robust security options to meet business governance requirements

The scale-out data lake enables organizations to lower costs while providing a rich repository of valuable insight with the potential to transform their businesses.

DATA LAKE WITH EMC AND PIVOTAL

EMC® Isilon™ scale-out storage, EMC DCA computing appliance and Pivotal software including Pivotal HD, Enterprise Hadoop Distribution, and HAWQ provide the foundation for a powerful and highly efficient scale-out data lake solution that combines massively scalable big data storage with robust computing and analytics capabilities. This solution allows you to create a scale-out data lake that allows you to gain new efficiencies and accelerate business insight.

Isilon is the only scale-out NAS platform with native support for the Hadoop Distributed File System (HDFS). With your unstructured data on Isilon, you can leverage Pivotal HD with HAWQ running on an EMC DCA to execute your batch and real-time data analytical queries without unnecessary capital expenditures, increased operational costs, and time-consuming replication of your Big Data to a separate infrastructure.

IN-PLACE HADOOP ANALYTICS WITH EMC AND PIVOTAL
EMC ISILON SCALE-OUT NAS

With Isilon scale-out storage solutions, you can have massive room for growth for your unstructured data assets. Isilon solutions scale easily from 16 terabytes (TB) to over 20 petabytes (PB) of capacity in a single cluster. By leveraging Isilon’s native HDFS support and in-place Big Data analytics capabilities, you can also avoid significant capital expenditures required for a separate, dedicated Hadoop infrastructure.

MULTI-PROTOCOL SUPPORT FOR OPERATIONAL FLEXIBILITY

INCREASE FLEXIBILITY

Along with native HDFS support, Isilon scale-out NAS storage include integrated support for a wide range of industry-standard protocols, including NFS, SMB, HTTP, FTP, and REST-based Object access for your cloud initiatives. As a result, with Isilon, your scale-out data lake allows you to simplify workflows, accelerate business analytics projects, support cloud initiatives, and get more value from your enterprise data.

REDUCE COSTS

With Isilon, you can streamline your storage infrastructure by consolidating large-scale file and unstructured data assets, eliminating silos of storage while providing the foundation for your scale-out data lake. Isilon allows you to achieve a storage utilization rate of over 80 percent. This unmatched efficiency means that you need less storage capacity and physical space to house the same amount of data than with other alternatives—reducing both initial capital outlay and ongoing costs. And because Isilon is simple to manage, it requires fewer IT resources for storage administration.

ACCELERATE RESULTS

Isilon’s in-place data analytics approach allows you to eliminate the time and resources required to replicate your Big Data into a separate infrastructure. For example, it can take over 24 hours to copy 100 TB of data over a 10Gb line. Instead, with Isilon, you can initiate data analytics projects immediately and get results in a matter of minutes. The ability to carry out in-place analytics without the lengthy data ingest phase enables support for analytics at the speed of business.

PROTECT AND SECURE YOUR BIG DATA ASSETS

EMC Isilon provides end-to-end data protection options for your unstructured data including fast and efficient data backup and disaster recovery. To help you meet compliance 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 secure access zones, to create isolated storage pools for specific departments within your organization.
The EMC Data Computing Appliance (DCA) is an integrated analytics platform that accelerates analysis of Big Data assets within a single integrated appliance. The DCA integrates Pivotal HD software with CPUs, memory and EMC Isilon scale-out storage to provide independent scaling of compute and storage resources. EMC DCA provides an ideal blend of flexibility, price and performance, to accelerate deployment of your Big Data analytics solution.

The EMC DCA, combined with Pivotal software including Pivotal HD and HAWQ, and EMC Isilon scale-out storage, provides a pre-configured, ready-to-run, enterprise-ready Apache Hadoop solution, reducing deployment time to a few hours rather than weeks or months as with traditional approaches.

**EXTREME ANALYTICS PERFORMANCE**

At the heart of the EMC Data Computing Appliance (DCA) is Pivotal HD, a fully-supported, enterprise-ready commercial distribution of Apache Hadoop. The core principle of the Pivotal DCA is to bring the processing power of SQL and MapReduce dramatically closer to the data, running analytics in parallel with data flowing efficiently between resources as needed. The result is industry-leading performance for big data analysis at an affordable TCO.

**INDUSTRY LEADING FLEXIBILITY AND SCALABILITY**

EMC DCAs are configured to match users’ needs, delivered ready to run and deployable within a few hours. Capabilities can be customized to the user’s needs, choosing from modules that support Hadoop, in-memory distributed systems and partner applications such as analytics, data visualization, machine learning, BI and ETL. Once deployed, DCA capacity can be scaled linearly by adding modules. New functions can also be added by adding new module types. DCAs can be configured from 1/4 rack to 11 racks supporting a vast range of configurations by mixing module types.

**COHERENT APPLIANCE WIDE ADMINISTRATION**

DCAs are easily administered regardless of configuration as all modules are managed and monitored through Pivotal Command Center, an appliance-wide administration tool. Integration with SNMP Network Management systems helps DCAs fit easily into most data center management frameworks.
PIVOTAL HD

- Maximize the value from all your data to enable the business.
- Leverage existing SQL skill sets for advanced analytics
- Accelerate time-to-value with an in-memory data grid for closed loop analytics

PIVOTAL HD

Pivotal HD offers the capabilities of Apache Hadoop, enhanced and packaged in a fully-supported, enterprise-ready distribution. In addition, Pivotal delivers value-added services with a rich, proven, parallel SQL query processing engine combined with a distributed in-memory system. Pivotal HD is the industry’s only closed loop batch and analytics processing for OLAP and OLTP with HDFS as the common data storage layer. By bringing analytics to bear on the day-to-day events of your business, you can build prescriptive systems that take the right actions in real-time.

PIVOTAL HD ARCHITECTURE

FAST, PROVEN SQL ANALYTICAL QUERY ENGINE FOR HADOOP

Unlike new SQL-on-Hadoop entrants, Pivotal HAWQ brings more than 10 years of innovation from the Greenplum database that has resulted in a rich, powerful SQL query optimizer and processor optimized to run analytical queries and mixed query workloads in massively parallel, distributed environments on top of HDFS. Because HAWQ is 100% SQL compliant, existing skill sets and products can be leveraged to accelerate data analytics projects dramatically, often with 100 times better performance than Hadoop alone.

HAWQ also has an external interface with HDFS, HBase and Hive, which allows you to read and query data stored in the Hadoop ecosystem and in addition load it directly into HAWQ. This also supports a wide range of data formats including Text, AVRO, and RCFile formats.

REAL-TIME DATA SERVICES FOR HADOOP

GemFire XD is an in-memory data grid that combines with Pivotal HD and HAWQ to provide the industry’s first production-quality platform for creating closed loop analytics solutions. GemFire XD enables the creation of low latency, scale out OLTP applications integrated out of the box with a big data store (HDFS). This provides sub-second response to applications, while allowing the data to be analyzed in the back end via HAWQ, Hive, or Map Reduce in real-time.
SUMMARY

A scale-out data lake can serve as a rich repository of valuable insight that can help you transform your business and gain significant advantages. EMC Isilon scale-out NAS, EMC DCA data computing appliance and Pivotal software, including Pivotal HD and HAWQ, provide the foundation for a powerful and highly efficient scale-out data lake solution that combines massively scalable big data storage, powerful computing and robust analytics capabilities.

TAKE THE NEXT STEP

Contact your EMC or Pivotal sales representative or authorized reseller to learn more about how a scale-out data lake can benefit your organization.

Also see our solutions in the EMC Store at https://store.emc.com/isilon.