



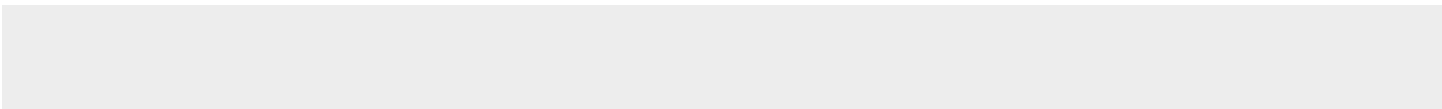
White Paper

---

## Applying the Cloud to Big Data Storage

# Contents

<b>Executive Summary</b>	<b>1</b>
<b>Storage Trends</b>	<b>2</b>
<i>Transitioning Storage Systems to Cloud Technologies</i>	2
<i>Commoditization of Storage Equipment</i>	3
<i>Move Towards Data Localization</i>	4
<b>Appistry CloudIQ Storage: Cloud Technology Applied to Big Data</b>	<b>6</b>
<i>Self-Organizing and Scalable</i>	6
<i>Geographically Aware</i>	7
<i>Reliable</i>	8
<i>Available</i>	9
<i>Manageable</i>	10
<i>Secure</i>	10
<b>Summary</b>	<b>11</b>



## Executive Summary

Traditional approaches to scaling storage and processing have begun to reach computational, operational and economic limits. These traditional approaches often depend on brittle clustering techniques and master/slave deployment patterns. While yesterday's applications have been well-served by these patterns, they fail to meet the needs of today's rapidly growing data sets.

In fact, any approach that depends on holding a collection of files, data elements or metadata within a few root, master or controller machines will eventually run out of space. Hierarchical models, which serve to organize or cluster the controller machines, can alleviate the pain but they put tremendous performance and reliability pressure on the root nodes. In addition, these models tend to require specialized hardware with a high dollars-per-terabyte cost, making large-scale deployment expensive.

Appistry believes that large-scale storage and processing requirements demand a cloud-based architectural approach. Since 2001, Appistry's efforts have focused on helping customers deliver mission-critical distributed systems. These experiences are readily applied to the challenges of so-called "big data" environments—systems where many terabytes or petabytes of data must be stored and processed. The challenges of large-scale storage and application environments are many, spanning factors such as scalability, capacity and performance to overall system cost and complexity.

In addition, as organizations become increasingly global and users demand low-latency access to their data, these challenges are intensified by geography and the laws of physics. Multiple storage centers must now hold copies of each file to maximize reliability and availability and minimize latency. Regional centers or edge locations may need to hold their own cache or repository of data. These physical boundaries must be hidden from users and administrators, who need to interact with data in multiple data centers but must view the entire distributed system as a single storage entity.

Cloud principles provide sound direction in addressing these challenges. In this paper we explore how these principles have paved the way for a new approach to delivering large-scale storage and application services.

We begin by looking at the key trends driving the need for a new approach to storage. We explore the ways in which cloud architectural principles are applied to create a new approach to data storage that is more appropriate for petabyte scale. Along the way, we discuss the key considerations in building real-life systems at this scale. Finally, Appistry CloudIQ Storage is introduced, and we show how it delivers each of the essential characteristics of a cloud storage system.

## Storage Trends

The following three technology trends are having a dramatic impact on the way big data challenges will be addressed:

- Transitioning Storage Systems to Cloud Technologies
- Commoditization of Storage Equipment
- Move Towards Data Localization

Industry progress in these areas provides solutions for the construction of large data storage systems.

### *Transitioning Storage Systems to Cloud Technologies*

Appistry is often engaged to help organizations address the challenges of petabyte-scale problems. As a result, we have gained a new understanding of how storage and computational systems should be constructed to achieve large scale. There are many drivers leading the industry away from traditional monolithic storage and computational approaches:

- **Cost.** Traditional NAS, SAN, and RDBMS solutions are too expensive at petabyte scale.
- **Complexity.** There is no single traditional storage solution that can simply scale to meet the needs of these users. Multiple systems must be joined with external management components to reach the needed capacity.
- **Management.** Complex systems have greatly increased management overhead and are difficult to maintain.
- **Reliability.** The reliability requirement for multi-petabyte systems are different. Because no single storage device can hold all of the necessary data, multiple systems need cooperate to provide reliability. Errors can often result in periods of data inaccessibility. In addition, multi-petabyte systems are typically architected to span multiple data centers. Traditional systems aren't constructed to handle these issues.

Cloud computing architectures, on the other hand, are characterized by their use of large quantities of affordable, commodity systems with directly-attached storage, all working in concert to provide the user with a single system view that transcends the performance and capacity of any single machine within the system. A storage system built in this manner provides the following attributes:

- **Scalability.** A cloud storage administrator is able to add additional computers, including their storage capacity, to a running system without a loss of availability of files or administrative functionality. Because tracking and membership are fully distributed and dynamic, the overall system can grow to tens of thousands of systems, or more.
- **Capacity.** The cloud storage system provides a global view or namespace, aggregating the capacity of all attached storage devices.
- **Reliability.** Cloud storage allows the user to specify how many copies of each file to maintain in the system. The cloud is aware of the loss of any machines in the system. When these errors occur, the cloud can alert the proper administrators and take appropriate action to recover the requested reliability level.
- **Geographic Distribution.** A single instance of a cloud storage system can be deployed across multiple data centers. The cloud is aware of the network topology and will mirror and distribute files across the network so that the loss of any one data center does not limit access to data.
- **Disaster Recovery.** The storage system is fully distributed; there is no central point of failure. Cloud storage can continue operation even when entire data centers have been removed from the system. Cloud storage also manages the merging of multiple data centers after a logical or physical separation occurs. Out-of-date files are located and reconciled without user-intervention whenever possible.
- **Availability.** Every computer in the cloud system is capable of serving access to files or administrative requests. Cloud storage is easily able to service a large number of client requests by distributing the work across many machines. The system is impervious to the loss of individual or even entire racks of machines.
- **Manageable.** Administrators are able to update computer configurations, system configurations, or update the cloud system itself without taking the files offline.
- **Heterogeneous.** Not all machines in the cloud system need to be constructed from similar hardware. The system needs to recognize the attributes of each attached computer and utilize their resources accordingly.

By taking a cloud-oriented approach to storage and compute, we are able to deliver a more powerful system. Moreover, because cloud storage systems are built with commodity components, they are much less expensive than traditional approaches.

### *Commoditization of Storage Equipment*

Historically, system architects and administrators have depended on increasingly larger and larger machines and devices to satisfy their growing computational and storage needs. These high-end, proprietary systems have come at a steep price in terms of both capital and operational costs, as well as in terms of agility and vendor lock-in. The advent of storage and computational systems based on cloud architectures results in an advantageous economic position for purchasers and users of these solutions.

**Fig 1. Cost of a One Petabyte Storage System**

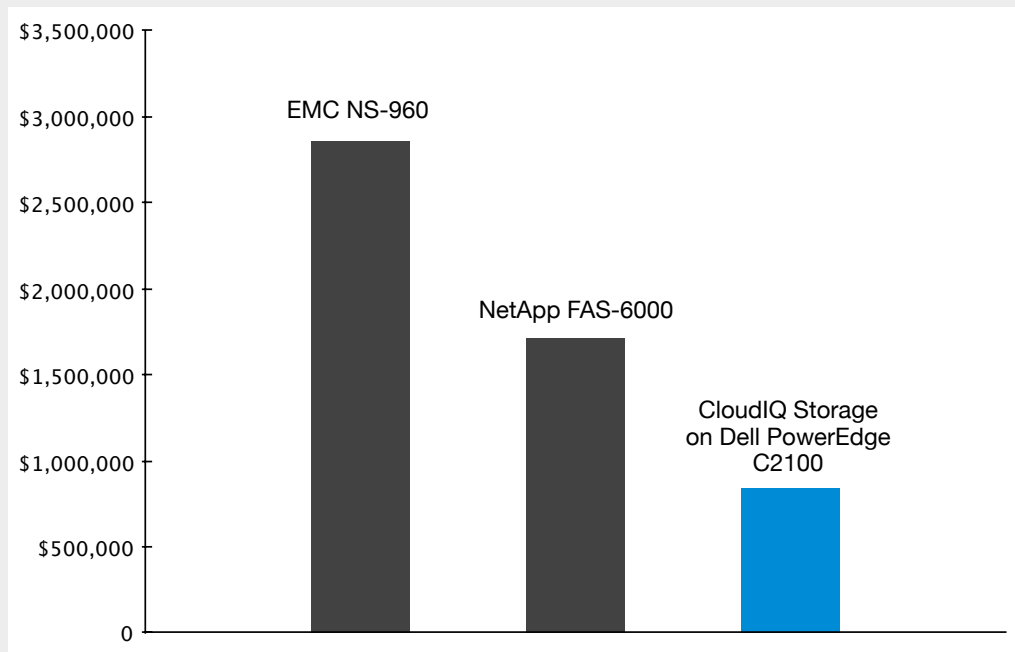


Figure 1 demonstrates the cost advantage of building a storage system from commodity components. Costs to build a one petabyte storage system using EMC Celerra and NetApp hardware are compared to those required to build a one petabyte system suitable for use with Appistry CloudIQ Storage<sup>1</sup>. The solution based on commodity servers is less than one-third of the cost of the EMC system and about one-half of the cost of the NetApp system. In addition, as we will see, the Appistry solution provides superior cloud storage features not readily attainable with traditional options.

### *Move Towards Data Localization*

In traditional system architectures, computational elements (i.e. application servers) and storage devices are partitioned into separate islands, or tiers. Applications pull data from storage devices via the local or storage area network, operate on it, and then push the results back to the storage devices. As a result, for most traditionally architected applications, the weak link in the system is the bottleneck between the application and its data.

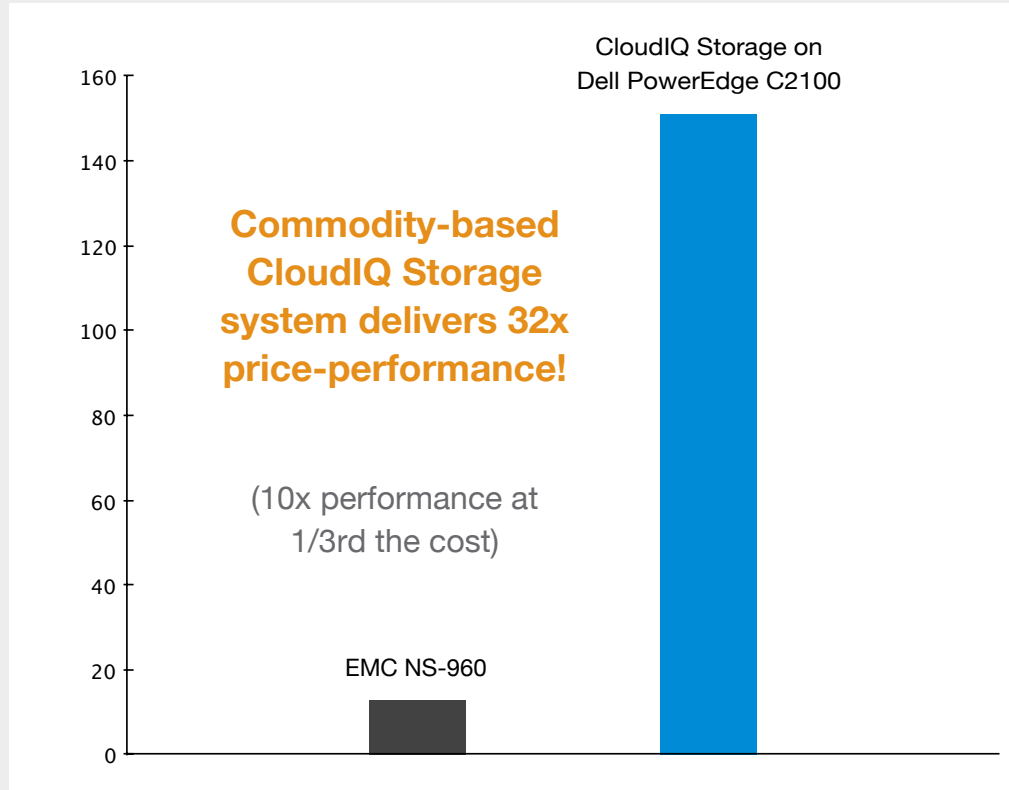
Data localization is the unification of storage devices with computational elements for the purposes of reducing computational latency and overcoming network bottlenecks. In a system exhibiting data locality, the work is moved to the data instead of the data being moved to the work.

CloudIQ Storage was built from the ground up to enable data localization, which Appistry calls *computational storage*. Other examples of data localization in practice

include the Apache Hadoop project – an implementation of the MapReduce algorithm initially popularized by Google, Netezza's data warehouse appliances, and various data caching technologies.

One way to compare the relative performance of traditional and cloud storage

**Fig 2. Relative Performance and Price-Performance**



approaches and to quantify the performance benefits of computational storage is to look at the aggregate bandwidth available between application processing and storage. In this example we use the same EMC Celerra as above, alongside a computational storage system based on Appistry CloudIQ Storage software and Dell hardware.

With the traditional architecture, all data must go through the SAN. The EMC Celerra streams data<sup>2</sup> out at a maximum of 16 GB/s (assuming ten 8 Gb Fibre Channel (FC) connections). In the computational storage case, data is streamed off of each drive at SATA bus speed (300 MB/s) and processed local to the server. As this is happening on each of 12 SATA ports on each of the 42 machines that comprise the system, the aggregate bandwidth is 151 GB/s.

As shown in Figure 2, this order-of-magnitude performance increase comes alongside

a savings of 70 percent. The net result is a 30x price-performance difference in favor of the computational storage system. This analysis is confirmed by real-world experience. Both Appistry and Netezza report performance improvements of 10-100x relative to traditional architectures for applications utilizing their technologies.

Taken together, the impact of cloud computing architectures, the commoditization of storage and compute, and the move towards data localization are revolutionizing the delivery of data-intensive applications; solving problems once thought to be unsolvable because of economic or temporal concerns has now become possible.

## **Appistry CloudIQ Storage: Cloud Technology Applied to Big Data**

Appistry CloudIQ Storage applies cloud computing architectural principles to create a scalable, reliable and highly cost-effective file storage system with no single points of failure, using only commodity servers and networking.

A CloudIQ Storage system is composed of multiple computers at one or more data centers, as depicted in Figure 3. CloudIQ Storage coordinates the activity of each of the computers and aggregates their attached storage to expose a single, logical data store to the user. The system is fully decentralized: each computer is a complete CloudIQ Storage system unto itself, but is aware of other members of the same storage system and shares storage responsibilities accordingly.

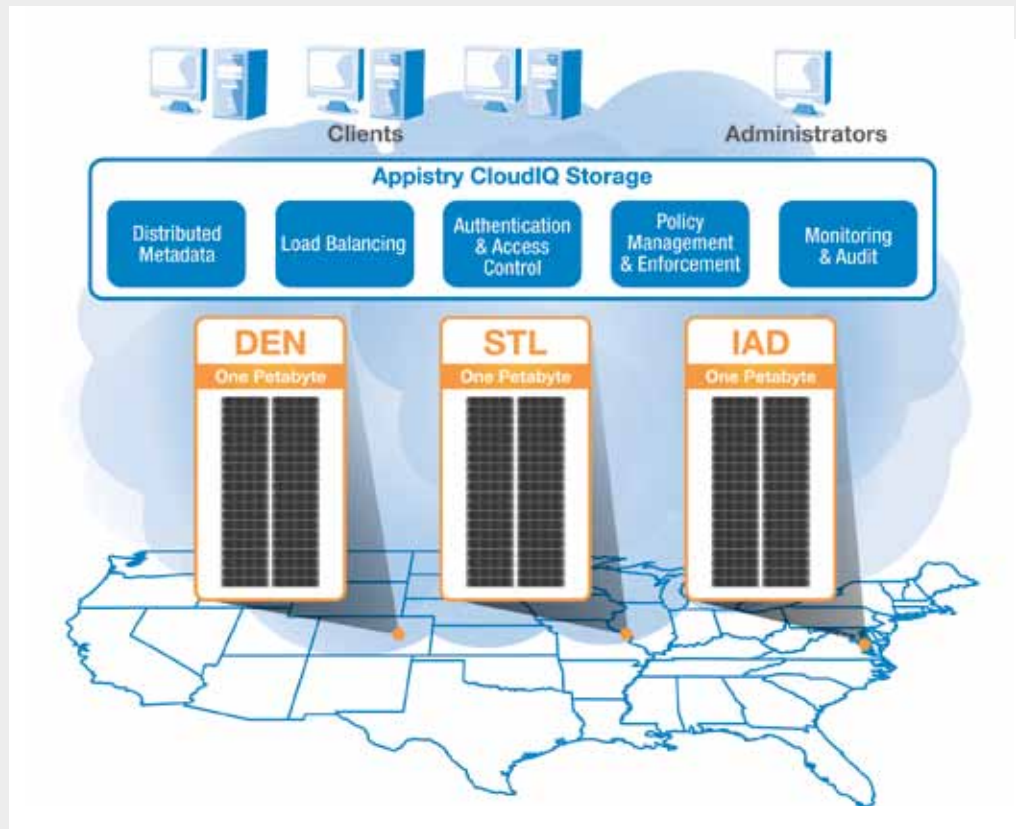
The remainder of this section describes several of the major functional characteristics of the CloudIQ Storage system. We consider these to be essential architectural characteristics of any cloud-based storage system.

### *Self-Organizing and Scalable*

Appistry believes that fully distributed systems with self-healing and self-organizing properties are the path to solving big data challenges. The foundation of the CloudIQ storage architecture is a lightweight, yet robust membership protocol. This protocol updates the member machines with the addition, removal or unexpected loss of computers dedicated to the storage system. This shared membership data contains enough information about the members of the cloud that each individual machine is capable of assessing its location and responsibilities. These responsibilities include establishing proper communication connections and responding to system events requiring healing actions. Even though there is no central control structure, the system is capable of self-organizing thousands of machines.

An administrator can easily add or remove machines or update configurations. The members of the cloud, acting independently, will share information quickly and reconfigure appropriately. The storage cloud can elastically scale up to handle multiple petabytes without heavy management overhead.

**Fig 3. CloudIQ Storage Creates a Unified Cloud Storage System Spanning Multiple Geographies.**



### *Geographically Aware*

One desired feature of a robust storage system is location awareness. Computers within the CloudIQ Storage environment can use location awareness to make informed decisions about reliability configurations and to optimize the handling of user requests. CloudIQ Storage introduces the notion of a territory to be a logical collection of machines classified as a group for the purpose of data distribution.

Users typically assign territories in one of several ways:

- **Computer Rack or Network Switch.** This configuration allows an administrator to instruct a storage cloud to distribute files and functionality across systems within a single data center.
- **Data Center.** This configuration allows an administrator to inform the cloud to distribute files and functionality between data centers.
- **User-Based.** For storage clouds that span multiple geographies, it is beneficial to

inform the system which computers are dedicated to individual user groups. Often this is a similar configuration to the data center option.

- **Hardware-Based.** This configuration allows different configurations of hardware to be grouped together. These groups provide the administrator with a method to store data on specialized hardware for different needs. For example, within a data center one might have two territories of low-latency computers set up across racks for availability. A third collection of machines might be constructed of higher-storage-density, higher-latency hardware to keep costs low while maintaining a third copy of the data.

Territory settings can be configured on a machine-by-machine basis. Administrators can choose from any of these use cases or develop hybrid configurations that meet their needs.

CloudIQ Storage uses territory settings to implement the behaviors described in the remainder of this section.

### *Reliable*

CloudIQ Storage provides high levels of reliability by distributing files and their associated metadata throughout the storage system. Each copy of a file possesses audit and configuration information needed to guarantee reliability requirements are achieved. The metadata of each file contains information on:

- **Reliability Needs.** How many copies of a file need to be maintained?
- **Territory Distribution.** Which territories can/should be used to keep a copy of the files?
- **Update History.** What is the version history of each file?

The reliability requirements of each file in the system are distributed across the machines in the CloudIQ Storage system. Each machine watches over a subset of files in the system. When these monitors detect system changes, the following actions occur to guarantee the reliability needs of the system:

- **File Reliability Checks.** Each monitor examines the files for which it is responsible. If a machine holding a copy of the file has been lost, additional copies of the file are created.
- **File Integrity Checks.** If a dormant or lost machine attempts to introduce an old copy of an existing file, the system reconciles the version against the metadata of the newer files and acts to reconcile the difference.
- **System Monitoring Reconfiguration.** As machines are introduced or lost, the responsibilities for watching files are adjusted for the new system configuration.

- **File Placement Reconfiguration.** As new machines become available, the monitors may decide to redistribute files. The reconfiguration distributes the storage needs and service requests more equally across machines in the storage cloud. Files may also need to be repositioned to meet territory placement requirements.

As the storage cloud grows and changes with new hardware, new network connections, and configuration changes, the cloud storage system will constantly maintain the proper file distribution and placement.

### *Available*

CloudIQ Storage provides extraordinary levels of availability due to the completely decentralized nature of the architecture. Every computer within the system is capable of serving file ingestion or file retrieval requests. Therefore, the total bandwidth in and out of the system is the aggregate of that of all of the machines participating in the cloud.

In addition, even though multiple copies of the file are present in the cloud storage system, the user gets only a single, logical view of the system. The CloudIQ Storage architecture resolves the multiple territories, copies and even versions of each file to deliver the requested file to the user.

When a file retrieval request arrives at a computer in a cloud storage system, several actions occur to provide the user with their data quickly:

- **File Location.** The computer servicing a file request locates which machines in the cloud hold the requested file using consistent hashing and a distributed hash tables. No single machine holds the entire file directory, as it would become a performance bottleneck or a point of failure. Lookups are a constant time operation that returns the machines within the system holding a copy of the file.
- **Machine Selection.** Once the target machines holding the file have been identified, the requesting machine can choose which machine is optimal for retrieving the file. This choice can be made based on factors such as network proximity and machine utilization.
- **File Retrieval.** Once the machine is selected, the file can be retrieved by the client.

In addition to optimized read operations, the cloud storage solution provides always writable semantics using a concept called eventual consistency. In an eventually consistent system, write operations always succeed as long as the system can access the number of nodes required by policy for a successful write (one, by default). During this write operation, audit information is stored with the file's metadata so that any additional copies or version reconciliation can be performed later. Eventually consistent systems are not right for every storage application, but it is ideal for write once, read many style systems.

The availability, reliability, and location awareness features of a cloud storage solution bring the highest level of disaster recovery available to a storage administrator. The system can lose a machine, a rack, or even an entire data center and the system remains capable of performing all necessary operations for the user.

### *Manageable*

Management and ease-of-use features are essential for the creation of a robust cloud storage system. When dealing with hundreds or thousands of machines, management operations must be simplified. CloudIQ Storage ensures this by providing the following attributes:

- **Always Available Operation.** Any configuration changes made to the system must not remove the availability of files. In the event that multiple machines need to be taken off line for updates, the system must have a strategy for keeping files available. This may be achieved using territories. If two territories hold copies of the same files, machines in one territory can temporarily be taken off-line for updates and the second territory can serve the files. Any file updates performed during the downtime will be automatically reconciled using the monitoring and reliability features of the cloud.
- **Configurable Reliability Settings.** Administrators can declare how many copies of a file should be stored in the storage cloud. A cloud-wide setting is established, which may be overridden on a file-by-file basis.
- **Real-Time Computer Injection.** When the storage system needs more capacity, the administrator needs to be able to add machines without affecting the availability of any file.
- **Real-Time Computer Decommissioning.** When it is decided that a computer is no longer required, the administrator needs operations to gracefully remove the computer from processing requests and move its files to other machines within the cloud.
- **Auditing.** Important operations, events, and system messages need to be saved.
- **System-Wide Configuration Changes.** All configuration changes need to propagate across all machines with a single operation.

Because management tasks in the storage cloud are virtualized and automated, a small number of system administrators can easily maintain a large number of computers storing petabytes of data.

## Secure

CloudIQ Storage implements a flexible security model designed to allow multitenant operation while ensuring the security of user data. Just as the physical storage cloud may be partitioned into territories, the logical storage cloud may be partitioned into “spaces,” with each space representing a hierarchical collection of files under common control. Access to each space, and to each file within a space, is governed by an access control list (ACL) that assigns access rights for users, groups and the file’s owner.

To facilitate secure operation of the cloud, administration rights to the system are divided into a series of distinct privileges that may be assigned to individual users or groups.

## Summary

Traditional storage and computational offerings fail to meet the needs of today’s big data environments. These approaches have been characterized by isolated pools of expensive storage and the constant movement of data from where it lives to the application servers and users who need it. Attempting to meet demanding capacity and performance requirements in this manner often necessitates the purchase of extremely costly, special-purpose hardware. Yet, local and wide-area network bottlenecks remain a challenge.

Petabyte-scale environments dictate the need for distributed, high-performing solutions that bring the work to the data, not the other way around. In this paper we have demonstrated how the cloud computing architectural approach provides the key to meeting the challenges posed by big data.

Appistry CloudIQ Storage is software that applies these principles to deliver robust private cloud storage environments. Storage clouds based on CloudIQ Storage exhibit essential characteristics we propose for any cloud storage system: individual resources self-organize without centralized control, yielding extreme scalability; the system spans data centers and optimizes behavior based on topology; high levels of reliability and availability are transparently ensured; system management is policy-based, automated and virtualized.

---

<sup>1</sup>EMC Celerra: \$2,860,000; NetApp FAS-600: \$1,714,000; Dell C2100: \$840,000. All prices quoted at list. EMC, NetApp price estimates courtesy Backblaze blog: <http://blog.backblaze.com/2009/09/01/petabytes-on-a-budget-how-to-build-cheap-cloud-storage/>

<sup>2</sup>See <http://www.emc.com/collateral/hardware/specification-sheet/h6035-celerra-ns960-ss.pdf>.

Copyright © 2010 -2011 Appistry, Inc.

*Appistry, CloudIQ and the Appistry logo are registered trademarks of Appistry, Inc. All rights reserved. Dell and PowerEdge are trademarks of Dell, Inc. All other registered and unregistered trademarks are the sole property of their respective owners.*

**Appistry** • 10845 Olive Blvd., Suite 260, St. Louis, MO 63141 • 314.336.5080 • [info@appistry.com](mailto:info@appistry.com)